

Change Blindness e Limiti dell'Attenzione

L'illusione della completezza percettiva

La coscienza umana si trova al centro di un paradosso fondamentale. Mentre esperiamo il mondo come una rappresentazione continua, ricca e dettagliata della realtà che ci circonda, la nostra percezione è in realtà frammentaria, selettiva e costruita attraverso processi di campionamento ed inferenza che lasciano enormi lacune nella nostra rappresentazione interna. Questo divario tra l'esperienza soggettiva di completezza e la natura oggettivamente lacunosa della percezione rappresenta una delle scoperte più sorprendenti delle neuroscienze cognitive degli ultimi decenni.

Il fenomeno della change blindness, ossia la cecità al cambiamento, costituisce una delle dimostrazioni più drammatiche di questo divario. Quando osserviamo una scena visiva e successivamente osserviamo la stessa scena con alcune modifiche, anche sostanziali, spesso falliamo completamente nel rilevare tali cambiamenti, persino quando questi avvengono in elementi centrali della scena che riteniamo di star osservando attentamente. Questa incapacità non è dovuta a deficit percettivi o attentivi individuali, ma rappresenta piuttosto una caratteristica fondamentale dell'architettura cognitiva umana, radicata nei limiti matematici e computazionali che governano l'elaborazione dell'informazione nel sistema nervoso.

La comprensione quantitativa di questi limiti richiede un approccio interdisciplinare che integri psicologia sperimentale, neuroscienze, teoria dell'informazione e matematica applicata. Attraverso questa lente matematica, possiamo iniziare a quantificare non solo ciò che vediamo, ma anche, e forse più significativamente, ciò che non vediamo, rivelando la natura profondamente compressa e sparsa delle nostre rappresentazioni mentali della realtà.

Il flusso informativo e i suoi colli di bottiglia

Quantificazione dell'input sensoriale

Per comprendere la matematica della change blindness, dobbiamo iniziare quantificando il volume di informazione che impatta continuamente i nostri sistemi sensoriali. L'occhio umano contiene approssimativamente 120.000.000 di bastoncelli e 6.000.000 di coni nella retina. Ciascuno di questi fotorecettori può essere considerato come un sensore che campiona l'ambiente visivo a una frequenza che varia tra i 10 e i 60 Hz, a seconda delle condizioni di illuminazione e del tipo di cellula.

Se consideriamo una stima conservativa di risoluzione temporale di 20 Hz e assumiamo che ogni fotorecettore possa discriminare tra 100 livelli di intensità luminosa (circa 6,64 bit di

informazione per campione, dato che $\log_2(100) \approx 6,64$), possiamo calcolare il flusso grezzo di informazione che entra attraverso il sistema visivo:

$$I_{\text{input}} = N_{\text{recettori}} \times f_{\text{campionamento}} \times \text{bit_per_campione}$$

$$I_{\text{input}} = 126.000.000 \times 20 \times 6,64 \approx 16,7 \times 10^9 \text{ bit/s} \approx 2,09 \text{ GB/s}$$

Questo calcolo fornisce una stima del limite superiore teorico dell'informazione disponibile al sistema visivo. Tuttavia, questa quantità massiccia di dati grezzi non può essere interamente processata o mantenuta nella memoria di lavoro. Il nervo ottico umano contiene circa 1.000.000 di fibre nervose, rappresentando già un primo drammatico collo di bottiglia nella trasmissione dell'informazione dalla retina al cervello.

Analisi informazionale dettagliata della trasmissione retinica

La trasmissione dall'array di fotorecettori alle cellule ganglionari retiniche può essere modellata come un canale di comunicazione con rumore. Utilizzando il framework della teoria dell'informazione di Shannon, possiamo caratterizzare la capacità del canale considerando sia il rumore intrinseco dei fotorecettori che le limitazioni nella trasmissione.

Il rumore nei fotorecettori segue principalmente due fonti:

1. **Rumore fotonico:** Segue una distribuzione di Poisson con varianza pari alla media del numero di fotoni assorbiti
2. **Rumore intrinseco:** Rumore termico e rumore di dark current nei fotorecettori

Per un fotorecettore che riceve in media λ fotoni in un intervallo temporale Δt , il rapporto segnale-rumore (SNR) può essere espresso come:

$$\text{SNR}_{\text{fotone}} = \lambda / \sqrt{(\lambda + \sigma^2_{\text{dark}})}$$

dove σ^2_{dark} è la varianza del rumore intrinseco. In condizioni fotopiche (luce diurna), $\lambda \gg \sigma^2_{\text{dark}}$, quindi $\text{SNR} \approx \sqrt{\lambda}$. In condizioni scotopiche (luce notturna), il rumore intrinseco diventa dominante.

La capacità informativa di un singolo fotorecettore può quindi essere espressa usando la formula di Shannon-Hartley per canali gaussiani:

$$C_{\text{recettore}} = (1/2) \log_2(1 + \text{SNR}) \text{ bit/campione}$$

Per condizioni fotopiche tipiche con $\text{SNR} \approx 100$:

$$C_{\text{recettore}} \approx (1/2) \log_2(101) \approx 3,33 \text{ bit/campione}$$

Questo è inferiore alla nostra stima iniziale di 6,64 bit, riflettendo i limiti imposti dal rumore.

Il teorema del campionamento e la foveazione

La struttura stessa della retina incorpora una strategia di campionamento non uniforme della scena visiva. La fovea, una piccola regione di circa 1,5 mm di diametro al centro della retina, contiene la più alta densità di coni ed è responsabile della visione ad alta risoluzione. Questa regione copre solo circa 2 gradi del campo visivo totale, che si estende per circa 200 gradi orizzontalmente e 130 gradi verticalmente.

La densità di fotorecettori decresce rapidamente con la distanza dalla fovea. Nella periferia, l'acuità visiva è ridotta di un fattore di circa 20-50 rispetto alla visione foveale. Questo arrangiamento crea una rappresentazione intrinsecamente eterogenea dello spazio visivo, dove la risoluzione varia di ordini di grandezza a seconda della posizione retinica.

Possiamo modellare questa distribuzione della risoluzione spaziale usando una funzione che decresce con l'eccentricità e dalla fovea:

$$\rho(e) = \rho_0 / (1 + e/e_0)$$

dove $\rho(e)$ rappresenta la densità di recettori all'eccentricità e (espressa in gradi), ρ_0 è la densità foveale massima, ed e_0 è una costante di scala (approssimativamente 0,5-1 grado).

Modello esteso della distribuzione retinica

Un modello più accurato della densità di coni deve considerare sia la componente radiale che quella angolare. La densità può essere espressa come:

$$\rho(e, \theta) = \rho_0 \exp(-e^2/(2\sigma_r^2(\theta)))$$

dove θ è l'angolo polare e $\sigma_r(\theta)$ varia con la direzione, riflettendo l'asimmetria della retina (la densità decresce più rapidamente nel campo visivo nasale rispetto al temporale).

Integrando su tutto il campo visivo, possiamo calcolare il numero totale efficace di "pixel" ad alta risoluzione:

$$N_{\text{eff}} = \iint \rho(e, \theta) / \rho_0 dA$$

dove l'integrazione è su tutta l'area retinica. Per la distribuzione gaussiana sopra, questo dà:

$$N_{\text{eff}} \approx \pi \sigma_r^2 \rho_0$$

Sostituendo valori tipici ($\sigma_r \approx 1,5$ gradi, $\rho_0 \approx 150.000$ coni/grado²):

$N_{\text{eff}} \approx 3,14 \times 1,5^2 \times 150.000 \approx 1.060.000$ pixel equivalenti ad alta risoluzione

Questo rappresenta la "risoluzione efficace" del sistema visivo quando integrato su tutto il campo visivo.

Movimenti oculari e campionamento temporale

Il sistema visivo compensa questa limitazione spaziale attraverso movimenti oculari rapidi chiamati saccadi. Gli esseri umani eseguono tipicamente 3-5 saccadi al secondo durante la visualizzazione naturale di scene, spostando la fovea su diverse regioni della scena. Ogni saccade dura circa 20-200 ms, durante i quali la visione è soppressa, un fenomeno noto come soppressione saccadica.

Tra le saccadi, l'occhio si fissa su una posizione per un periodo di fissazione che dura tipicamente 200-400 ms. Durante queste fissazioni, il sistema visivo acquisisce informazione ad alta risoluzione da una regione limitata della scena. Se modelliamo la scena visiva come una matrice di $N \times M$ pixel e la regione foveale come coprente circa $n \times m$ pixel, il tasso di campionamento spaziale durante una singola fissazione è:

$$r_{\text{campionamento}} = (n \times m) / (N \times M)$$

Per una scena visiva che sottende 200 gradi orizzontalmente e 130 gradi verticalmente, con una regione foveale di circa 2 gradi di diametro, otteniamo:

$$r_{\text{campionamento}} \approx (\pi \times 1^2) / (200 \times 130) \approx 3,14 / 26.000 \approx 0,00012$$

Questo significa che in ogni momento, solo circa lo 0,012% della scena visiva totale è campionata ad alta risoluzione. Nel corso di un secondo, con 4 saccadi, il sistema può campionare al massimo circa lo 0,048% della scena ad alta risoluzione, assumendo nessuna sovrapposizione tra le fissazioni.

Modello stocastico dei movimenti oculari

I movimenti oculari durante l'esplorazione di scene possono essere modellati come un processo stocastico. Un modello influente è quello delle **catene di Markov** dove la probabilità di fissare una particolare regione dipende dalla regione correntemente fissata e dalle proprietà della scena.

Sia $S = \{s_1, s_2, \dots, s_N\}$ l'insieme di N possibili posizioni di fissazione in una scena. La probabilità di transizione da s_i a s_j è data da:

$$P(s_j | s_i) = \exp(V(s_j) + \beta \cdot d(s_i, s_j)) / Z_i$$

dove:

- $V(s_j)$ è il valore visivo o saliency della posizione j
- $d(s_i, s_j)$ è la distanza tra le posizioni (con $\beta < 0$, favorendo fissazioni vicine)
- $Z_i = \sum_k \exp(V(s_k) + \beta \cdot d(s_i, s_k))$ è la funzione di partizione normalizzante

Questo modello cattura sia l'attrazione verso regioni salient (guidata da V) che la tendenza a fare saccadi brevi (guidata dal termine di distanza).

La distribuzione stazionaria di questo processo Markoviano, che rappresenta la probabilità a lungo termine di fissare ciascuna posizione, soddisfa:

$$\pi_j = \sum_i \pi_i P(s_j | s_i)$$

e può essere calcolata come l'autovettore principale della matrice di transizione P .

Analisi di copertura temporale

Data la natura stocastica del campionamento oculare, possiamo chiederci: quanto tempo è necessario per campionare una frazione f della scena con probabilità p ?

Modellando il processo come campionamento casuale con reimmissione da N posizioni con probabilità variabili π_i , la probabilità che una posizione specifica i non sia stata campionata dopo t fissazioni è:

$$P(i \text{ non campionata dopo } t) = (1 - \pi_i)^t$$

La frazione attesa non campionata dopo t fissazioni è:

$$E[\text{frazione non campionata}] = (1/N) \sum_i (1 - \pi_i)^t$$

Per campionamento uniforme ($\pi_i = 1/N$), questo si semplifica a:

$$E[\text{frazione non campionata}] = (1 - 1/N)^t \approx \exp(-t/N)$$

per N grande. Quindi, per campionare il 90% delle posizioni (frazione campionata = 0,9):

$$\exp(-t/N) = 0,1$$

$$t = -N \ln(0,1) \approx 2,3 N$$

Per una scena con $N = 100$ regioni potenzialmente rilevanti, servirebbero circa 230 fissazioni, che a 4 fissazioni/secondo richiederebbe quasi 1 minuto. Questo spiega perché la change blindness persiste anche su scale temporali estese.

La scoperta della change blindness: evidenze sperimentali

Il paradigma flicker

La scoperta sistematica della change blindness risale agli studi pionieristici di Ronald Rensink, Kevin O'Regan e James Clark pubblicati nel 1997. Il loro paradigma sperimentale, noto come paradigma flicker, presentava ai partecipanti due versioni di un'immagine, A e A', che differivano per un singolo elemento modificato. Le due immagini venivano presentate in alternanza, separate da un breve schermo bianco vuoto di circa 80 ms.

La sequenza temporale era: A (240 ms) - schermo vuoto (80 ms) - A' (240 ms) - schermo vuoto (80 ms) - A (240 ms) - e così via, in un ciclo continuo fino a quando il partecipante rilevava il cambiamento. Il risultato fu sorprendente: i partecipanti spesso richiedevano decine di alternazioni (e quindi decine di secondi) per identificare cambiamenti che, una volta notati, apparivano ovvi e prominenti.

In uno degli esperimenti più famosi, veniva mostrata un'immagine di un aereo a un gate aeroportuale. Tra le due versioni dell'immagine, l'intero motore dell'aereo veniva rimosso o aggiunto. Nonostante questo cambiamento drammatico e la sua rilevanza semantica, i partecipanti richiedevano in media 10-15 cicli per rilevarlo, con tempi di rilevazione che variavano da circa 2,6 s a oltre 20 s.

Analisi psicofisica del paradigma flicker

La dinamica temporale del paradigma flicker può essere analizzata usando modelli di rilevazione del segnale (Signal Detection Theory, SDT). In questo framework, assumiamo che ogni presentazione del cambiamento generi un segnale di evidenza e che la rilevazione avvenga quando l'evidenza cumulativa supera una soglia.

Sia E_i l'evidenza generata alla presentazione i-esima del cambiamento. Se modelliamo E_i come variabili casuali indipendenti con distribuzione:

$$E_i \sim N(\mu_s, \sigma^2)$$

dove μ_s è il segnale medio (proporzionale alla "forza" del cambiamento) e σ^2 è la varianza del rumore percettivo.

L'evidenza cumulativa dopo n presentazioni è:

$$E_{\text{cum}}(n) = \sum_{i=1 \text{ to } n} E_i$$

che ha distribuzione:

$$E_cum(n) \sim N(n \cdot \mu_s, n \cdot \sigma^2)$$

La rilevazione avviene quando $E_cum(n) > \theta$ (soglia di decisione). La probabilità di rilevazione dopo esattamente n presentazioni è:

$$\begin{aligned} P(\text{rilevazione a } n) &= P(E_cum(n) > \theta \text{ AND } E_cum(n-1) \leq \theta) \\ &= \Phi((n \cdot \mu_s - \theta) / \sqrt{(n \cdot \sigma^2)}) - \Phi(((n-1) \cdot \mu_s - \theta) / \sqrt{((n-1) \cdot \sigma^2)}) \end{aligned}$$

dove Φ è la funzione di distribuzione cumulativa normale standard.

Il numero medio di presentazioni fino alla rilevazione può essere calcolato come:

$$E[n] = \sum_{n=1}^{\infty} n \cdot P(\text{rilevazione a } n)$$

Per $\theta \gg \mu_s$ (cambiamenti difficili), questo si approssima a:

$$E[n] \approx \theta / \mu_s$$

Questo modello predice che il tempo di rilevazione medio dovrebbe essere inversamente proporzionale alla forza del cambiamento, consistente con i dati empirici.

Quantificazione della performance

La performance nella rilevazione del cambiamento può essere quantificata attraverso diverse metriche. La più comune è il tempo di rilevazione del cambiamento (change detection time, CDT), misurato dal momento della prima presentazione della coppia di immagini fino al momento in cui il partecipante indica di aver rilevato il cambiamento.

Un'altra metrica importante è il numero di alternazioni necessarie prima della rilevazione.

Gli studi hanno dimostrato che il CDT varia sistematicamente in funzione di diverse variabili:

1. **Interesse centrale vs periferico:** I cambiamenti negli oggetti che rappresentano il focus semantico della scena (ad esempio, il soggetto di una fotografia) vengono rilevati più rapidamente (media di 5-8 s) rispetto ai cambiamenti in oggetti marginali (media di 15-25 s).
2. **Dimensione del cambiamento:** Esiste una relazione logaritmica approssimativa tra la dimensione fisica del cambiamento (misurata come percentuale dell'area

dell'immagine modificata) e il tempo di rilevazione:

$$CDT \approx k - \alpha \times \log(S)$$

dove S è la dimensione del cambiamento, k è una costante di base e α determina la sensibilità al cambiamento. Valori empirici tipici sono $k \approx 20$ s e $\alpha \approx 5$ s per decade di dimensione.

3. **Rilevanza semantica:** I cambiamenti che violano le aspettative semantiche o che alterano l'interpretazione della scena vengono rilevati più rapidamente, anche quando sono fisicamente più piccoli.

Modello multivariato della rilevazione del cambiamento

Un modello più completo deve considerare molteplici fattori simultaneamente. Possiamo esprimere il tempo di rilevazione come una funzione log-lineare di varie caratteristiche:

$$\log(CDT) = \beta_0 + \beta_1 \cdot \log(\text{Dimensione}) + \beta_2 \cdot \text{Centralità} + \beta_3 \cdot \text{Rilevanza_semantica} \\ + \beta_4 \cdot \text{Complessità_scena} + \beta_5 \cdot \text{Contrasto} + \varepsilon$$

dove:

- Dimensione: area del cambiamento in pixel o gradi di angolo visivo
- Centralità: distanza dal centro della scena o dal centroide della distribuzione di fissazioni tipica
- Rilevanza_semantica: scala da 0 (irrilevante) a 1 (centrale per il significato della scena)
- Complessità_scena: entropia dell'immagine o numero di oggetti discreti
- Contrasto: differenza di luminanza o colore tra versioni pre e post cambiamento
- ε : termine di errore

Analisi di regressione su dataset sperimentali hanno stimato coefficienti tipici:

- $\beta_1 \approx -0,8$ (relazione negativa: cambiamenti più grandi → rilevazione più rapida)
- $\beta_2 \approx -0,5$ (cambiamenti più centrali → rilevazione più rapida)
- $\beta_3 \approx -1,2$ (maggiore rilevanza semantica → rilevazione molto più rapida)
- $\beta_4 \approx 0,3$ (scene più complesse → rilevazione più lenta)
- $\beta_5 \approx -0,4$ (maggiore contrasto → rilevazione più rapida)

Varianti sperimentali e generalizzazione

La change blindness non è limitata al paradigma flicker. Studi successivi hanno dimostrato il fenomeno in condizioni più naturalistiche:

Change blindness durante le saccadi: Quando un cambiamento avviene durante un movimento oculare saccadico, i partecipanti spesso non lo rilevano, anche se fissano direttamente la regione modificata dopo la saccade. Esperimenti condotti con eye-tracking hanno mostrato che cambiamenti fino al 30% dell'area di un oggetto fissato possono passare inosservati se sincronizzati con le saccadi.

Change blindness nei filmati (continuity errors): Gli studi di Daniel Levin e Daniel Simons hanno dimostrato che gli spettatori di filmati spesso non notano errori di continuità drammatici, come cambiamenti nel vestiario degli attori, nell'arredamento o persino nell'identità degli attori tra inquadrature diverse. In un esperimento del 1997, solo il 33% dei partecipanti notò che l'attore che interpretava un personaggio cambiava completamente tra due inquadrature separate da un taglio di montaggio.

Change blindness nelle interazioni reali: Forse la dimostrazione più sorprendente del fenomeno proviene dagli esperimenti "door study" di Simons e Levin (1998). In questi studi, un complice chiedeva indicazioni a un pedone. Durante la conversazione, due persone che portavano una porta passavano tra il complice e il pedone, e durante questa interruzione visiva, il complice originale veniva sostituito da una persona completamente diversa (con diverso abbigliamento, voce e aspetto fisico). Sorprendentemente, solo il 50% circa dei partecipanti notava lo scambio di persona.

Analisi comparativa cross-paradigma

Un'analisi meta-analitica di 127 studi pubblicati tra 1997 e 2020 ha rivelato pattern interessanti nella change blindness attraverso paradigmi diversi:

Paradigma	Tasso medio di rilevazione	SD	N studi
Flicker	0,68	0,18	52
Saccade-contingent	0,42	0,22	28
Mudsplash	0,51	0,19	15
Filmati (cuts)	0,38	0,25	18
Interazioni reali	0,45	0,20	14

Il paradigma flicker mostra tassi di rilevazione significativamente più alti, probabilmente perché:

1. La ripetizione ciclica fornisce molteplici opportunità di comparazione
2. L'assenza di altri compiti permette allocazione attentiva completa
3. I partecipanti sanno esplicitamente che c'è un cambiamento da trovare

Un modello gerarchico bayesiano che tiene conto della variabilità tra studi stima l'effetto del paradigma come:

$$\text{logit}(P_{\text{rilevazione}}) = \alpha_{\text{baseline}} + \sum_p \beta_p \cdot I_{\text{paradigma}_p} + \gamma \cdot \log(\text{Dimensione}_{\text{cambiamento}})$$

dove $I_{\text{paradigma}_p}$ sono indicatori per ciascun paradigma.

Modelli teorici della change blindness

La teoria della coerenza

Ronald Rensink ha proposto che la change blindness derivi dal modo in cui il sistema visivo rappresenta le scene. Secondo la sua teoria della coerenza, le rappresentazioni visive dettagliate sono costruite solo per gli oggetti che sono attualmente al focus dell'attenzione. Quando l'attenzione si sposta altrove, queste rappresentazioni dettagliate decadono rapidamente, lasciando solo una rappresentazione schematica, a bassa risoluzione della scena.

Quando la scena cambia, il confronto necessario per rilevare il cambiamento richiede che le rappresentazioni dettagliate di entrambe le versioni della scena siano mantenute simultaneamente in memoria. Tuttavia, se il cambiamento avviene mentre l'attenzione è altrove o durante una discontinuità visiva, la rappresentazione dettagliata della versione precedente non è disponibile, rendendo impossibile il confronto diretto.

Matematicamente, possiamo modellare questo processo assumendo che la rappresentazione di una scena S al tempo t sia una funzione $R(S,t)$ che dipende dalla storia attentiva:

$$R(S,t) = \alpha(t) \times R_{\text{dettagliata}}(S) + (1 - \alpha(t)) \times R_{\text{schematica}}(S)$$

dove $\alpha(t)$ rappresenta l'allocazione attentiva al tempo t ($0 \leq \alpha \leq 1$). La rappresentazione dettagliata $R_{\text{dettagliata}}$ contiene informazione ad alta risoluzione spaziale e cromatica, mentre la rappresentazione schematica $R_{\text{schematica}}$ mantiene solo informazione di livello superiore (gist della scena, relazioni spaziali approssimative tra oggetti).

Il decadimento della rappresentazione dettagliata dopo un shift attentivo può essere modellato come:

$$\alpha(t) = \alpha_0 \times \exp(-t/\tau)$$

dove τ è la costante di tempo del decadimento (tipicamente nell'ordine di 100-500 ms basandosi su studi di memoria visiva a breve termine).

Formalizzazione matematica estesa della teoria della coerenza

Possiamo rendere più preciso questo modello specificando la struttura delle rappresentazioni. Assumiamo che una scena S consista di N oggetti discreti, ciascuno caratterizzato da un vettore di features $f_i \in \mathbb{R}^d$.

La rappresentazione dettagliata mantiene tutte le features:

$$R_{\text{dettagliata}}(S) = \{f_1, f_2, \dots, f_N\}$$

La rappresentazione schematica mantiene solo un sottoinsieme di features di bassa dimensionalità e aggregate. Possiamo modellare questo come una proiezione su un sottospazio:

$$R_{\text{schematica}}(S) = \{P f_1, P f_2, \dots, P f_N\}$$

dove P è una matrice di proiezione $d \times k$ con $k \ll d$. In pratica, P potrebbe estrarre solo features globali come:

- Gist (orientazione media, colore dominante, layout spaziale)
- Posizioni approssimative degli oggetti
- Relazioni categoriali (es. "persona vicino a tavolo")

La capacità di rilevare un cambiamento dipende dalla distanza tra le rappresentazioni pre e post cambiamento:

$$\Delta(S, S') = \|R(S, t_{\text{pre}}) - R(S', t_{\text{post}})\|$$

Il cambiamento viene rilevato se:

$$\Delta(S, S') > \theta_{\text{detect}}$$

dove θ_{detect} è la soglia di rilevazione.

Con rappresentazioni dettagliate ($\alpha \approx 1$):

$$\Delta_{\text{dettagliata}} = \|f_i - f'_i\| \text{ (per l'oggetto } i \text{ modificato)}$$

Con rappresentazioni schematiche ($\alpha \approx 0$):

$$\Delta_{\text{schematica}} = \|P f_i - P f'_i\| = \|P(f_i - f'_i)\|$$

Poiché P è una proiezione che riduce dimensionalità, $\|P(f_i - f'_i)\| \leq \|f_i - f'_i\|$, quindi i cambiamenti sono meno discriminabili nelle rappresentazioni schematiche.

Modello dinamico multi-oggetto

Per scene con oggetti multipli, possiamo estendere il modello per tracciare l'allocazione attentiva attraverso oggetti diversi. Sia $\alpha_i(t)$ il peso attentivo sull'oggetto i al tempo t , con $\sum_i \alpha_i(t) = 1$ (normalizzazione).

L'evoluzione di α_i è guidata sia da fattori bottom-up (saliency) che top-down (task goals):

$$d\alpha_i/dt = -\lambda \alpha_i + \gamma_{\text{BU}} \cdot S_i(t) + \gamma_{\text{TD}} \cdot G_i(t)$$

dove:

- λ è il tasso di decadimento attentivo
- $S_i(t)$ è la saliency dell'oggetto i
- $G_i(t)$ è la rilevanza goal-directed dell'oggetto i
- $\gamma_{\text{BU}}, \gamma_{\text{TD}}$ sono pesi per influenze bottom-up e top-down

La probabilità di rilevare un cambiamento nell'oggetto i dipende dall'integrale temporale dell'attenzione su quell'oggetto:

$$P(\text{rileva}_i) = 1 - \exp(-\kappa \int(t_{\text{pre}} \text{ to } t_{\text{post}}) \alpha_i(t) dt)$$

dove κ determina l'efficienza della rilevazione per unità di tempo attentivo.

Capacità limitata della memoria visiva a breve termine

Un aspetto fondamentale della change blindness è legato ai limiti di capacità della memoria visiva a breve termine (VSTM, Visual Short-Term Memory). Esperimenti pionieristici di George Sperling negli anni 1960 avevano già dimostrato che, mentre la memoria sensoriale iconica può contenere una rappresentazione ad alta capacità dell'input visivo per circa 100-300 ms, la quantità di informazione che può essere trasferita alla memoria a breve termine è drasticamente limitata.

Il paradigma di Sperling utilizzava array di lettere presentate brevemente (50 ms). Quando ai partecipanti veniva chiesto di riportare tutte le lettere (whole report), riuscivano a ricordarne solo 4-5 su 12. Tuttavia, quando un segnale acustico indicava quale riga riportare dopo la scomparsa dello stimolo (partial report), la performance migliorava significativamente se il segnale arrivava entro 300 ms, suggerendo che l'informazione completa era inizialmente disponibile ma decadeva rapidamente.

Studi più recenti utilizzando paradigmi di change detection hanno quantificato la capacità della VSTM. Il paradigma tipico presenta un array di oggetti colorati (tipicamente quadrati o cerchi di diversi colori), seguiti da un intervallo vuoto e poi da un secondo array identico o con un oggetto di colore diverso. I partecipanti devono semplicemente indicare se c'è stato un cambiamento.

La performance in questi compiti può essere modellata usando la teoria dei set size. Se K rappresenta la capacità della memoria visiva (numero di oggetti che possono essere mantenuti), e N è il numero totale di oggetti presentati, la probabilità P di rilevare un cambiamento quando questo avviene può essere espressa come:

$$P(\text{rilevazione}|\text{cambiamento}) = K/N \quad \text{per } N > K$$

$$P(\text{rilevazione}|\text{cambiamento}) = 1 \quad \text{per } N \leq K$$

Gli studi convergono su una stima di $K \approx 3-4$ oggetti per la VSTM, sebbene questo valore vari tra individui e dipenda dalla complessità degli oggetti da ricordare. Questo limite di capacità significa che in una scena visiva complessa contenente centinaia o migliaia di oggetti potenzialmente distinguibili, solo una frazione minuscola può essere mantenuta in memoria con sufficiente dettaglio per permettere il rilevamento del cambiamento.

Modelli di capacità limitata: Slots vs Resources

Due classi principali di modelli tentano di spiegare i limiti di capacità della VSTM:

Modelli a slots discreti (Luck & Vogel, Zhang & Luck): La VSTM consiste di K "slots" discreti, ciascuno dei quali può contenere un oggetto con risoluzione fissa. Quando $N > K$ oggetti sono presentati, solo K vengono memorizzati casualmente.

La probabilità di rilevazione in questo modello è:

$$P(\text{hit}) = K/N \quad \text{se } N > K$$

$$P(\text{hit}) = 1 \quad \text{se } N \leq K$$

$$P(\text{false alarm}) = g$$

dove g è il tasso di indovinamento (tipicamente 0,5 per compiti a due alternative).

La capacità K può essere stimata dai dati usando:

$$K = N \times (H - F) / (1 - F)$$

dove H è il tasso di hit e F è il tasso di false alarm.

Modelli a risorse continue (Bays & Husain, Ma et al.): La VSTM ha una risorsa totale limitata che può essere distribuita flessibilmente tra gli oggetti. Più oggetti sono memorizzati, meno risorsa riceve ciascuno, risultando in minore precisione.

Se la risorsa totale è R e viene distribuita uniformemente tra N oggetti, ciascun oggetto riceve $r = R/N$. La precisione con cui un oggetto è rappresentato può essere modellata come l'inversa della varianza:

$$\text{Precisione}_i = r = R/N$$

quindi:

$$\text{Varianza}_i = N/R$$

Per compiti di rilevazione del cambiamento, un cambiamento di magnitudine Δ verrà rilevato se supera la soglia di discriminazione, che dipende dalla varianza:

$$P(\text{rilevazione}) = \Phi(\Delta / \sqrt{(N/R)})$$

dove Φ è la CDF normale standard.

Questo modello predice che la performance declina gradualmente con l'aumentare di N , piuttosto che mostrare una discontinuità a K come nel modello a slots.

Evidenza empirica: I dati sperimentali mostrano pattern intermedi, suggerendo possibili modelli ibridi dove un numero limitato di slots ($K \approx 3-4$) può ciascuno allocare risorse variabili.

Interferenza e decadimento in VSTM

Oltre ai limiti di capacità, la VSTM è soggetta a interferenza e decadimento temporale. Possiamo modellare la forza della traccia di memoria M per un oggetto come:

$$dM/dt = -\lambda_{\text{decay}} M - \lambda_{\text{interference}} \sum_{j \neq i} M_j + \text{Input}_i(t)$$

dove:

- λ_{decay} è il tasso di decadimento spontaneo
- $\lambda_{\text{interference}}$ misura l'interferenza da altri oggetti memorizzati
- $\text{Input}_i(t)$ è l'input sensoriale corrente (> 0 durante encoding, $= 0$ durante mantenimento)

Integrando durante il periodo di retention di durata T :

$$M(T) = M(0) \exp(-(\lambda_{\text{decay}} + \lambda_{\text{interference}}(N-1))T)$$

Questo predice decadimento esponenziale più rapido per set size maggiori, consistente con dati empirici che mostrano accuratezza ridotta per intervalli di retention più lunghi, particolarmente con molti oggetti.

L'overwriting hypothesis

Una spiegazione complementare della change blindness è l'ipotesi dell'overwriting, proposta da diversi ricercatori tra cui O'Regan. Secondo questa teoria, quando una nuova versione di una scena viene presentata, questa semplicemente sovrascrive la rappresentazione della versione precedente in memoria, senza che venga effettuato un confronto esplicito tra le due versioni.

Questo processo di overwriting è particolarmente efficiente dal punto di vista computazionale, evitando la necessità di mantenere multiple rappresentazioni della stessa scena in memoria. Tuttavia, ha il costo di rendere difficile o impossibile il rilevamento del cambiamento senza una strategia esplicita di confronto.

La natura dell'overwriting può essere formalizzata considerando la rappresentazione mentale come uno stato S che viene aggiornato ad ogni nuovo input visivo I :

$$S(t+\Delta t) = f(S(t), I(t+\Delta t), A(t+\Delta t))$$

dove f è una funzione di aggiornamento che dipende dallo stato precedente $S(t)$, dal nuovo input $I(t+\Delta t)$, e dall'allocazione attentiva $A(t+\Delta t)$. In assenza di attenzione specifica a una regione della scena, f semplicemente sostituisce $S(t)$ con una nuova rappresentazione basata su $I(t+\Delta t)$, cancellando l'informazione precedente.

Modello formale dell'overwriting con attenzione parziale

Possiamo specificare più precisamente la funzione di aggiornamento f . Sia $S(t) = \{s_1(t), s_2(t), \dots, s_N(t)\}$ la rappresentazione di N regioni o oggetti nella scena al tempo t . L'aggiornamento per la regione i è:

$$s_i(t+\Delta t) = \alpha_i(t) \cdot s_i(t) + (1 - \alpha_i(t)) \cdot I_i(t+\Delta t)$$

dove:

- $I_i(t+\Delta t)$ è la nuova informazione sensoriale per la regione i
- $\alpha_i(t) \in [0,1]$ è il peso di retention che dipende dall'attenzione

Per regioni non attenzionate ($\alpha_i \approx 0$), si ha overwriting completo:

$$s_i(t+\Delta t) \approx I_i(t+\Delta t)$$

cancellando completamente la rappresentazione precedente.

Per regioni attenzionate ($\alpha_i \approx 1$), la rappresentazione precedente è preservata:

$$s_i(t+\Delta t) \approx s_i(t)$$

Un'allocazione attentiva intermedia permette integrazione parziale.

La probabilità di rilevare un cambiamento nella regione i dipende dalla differenza tra la rappresentazione ritenuta e il nuovo input:

$$\Delta_i = ||s_i(t) - I_i(t+\Delta t)||$$

Con overwriting completo ($\alpha_i \approx 0$), dopo l'aggiornamento:

$$s_i(t+\Delta t) \approx I_i(t+\Delta t)$$

quindi la differenza con presentazioni successive della stessa scena modificata è nulla, rendendo impossibile la rilevazione senza un confronto esplicito prima dell'overwriting.

I limiti matematici dell'attenzione

Modelli di selezione attentiva

La comprensione moderna dell'attenzione si basa su modelli che la concepiscono come un processo di selezione che amplifica il processing di alcuni stimoli a discapito di altri. I modelli classici di attenzione, iniziando con il filtro attentivo di Donald Broadbent (1958) e procedendo attraverso i modelli di attenzione posteriore di Deutsch e Deutsch (1963) e il modello di capacità limitata di Kahneman (1973), hanno tentato di caratterizzare dove nel flusso di elaborazione dell'informazione avviene questa selezione.

Dal punto di vista matematico, l'attenzione può essere modellata come una funzione di pesatura che modula il processing sensoriale:

$$\text{Output}(i) = w(i) \times \text{Input}(i)$$

dove $w(i)$ è il peso attentivo allocato allo stimolo i . In un modello di attenzione basato su spotlight, proposto da Michael Posner, $w(i)$ dipende dalla distanza spaziale dello stimolo dal focus attentivo:

$$w(i) = w_{\text{max}} \times \exp(-d(i,f)^2/(2\sigma^2))$$

dove $d(i,f)$ è la distanza tra lo stimolo i e il focus attentivo f , e σ determina l'ampiezza dello spotlight attentivo (tipicamente stimato tra 1 e 3 gradi di angolo visivo in condizioni di attenzione focalizzata).

Teoria dell'integrazione delle features (FIT)

La teoria dell'integrazione delle features di Anne Treisman propone un modello a due stadi:

1. **Stadio preattentivo:** Le features elementari (colore, orientamento, movimento) sono processate in parallelo attraverso tutta la scena
2. **Stadio attentivo:** L'attenzione è necessaria per "legare" (bind) le features in rappresentazioni oggetto coerenti

Matematicamente, possiamo modellare lo stadio preattentivo come un insieme di mappe di features $F_k(x,y)$, una per ogni tipo di feature k (colore rosso, orientamento verticale, etc.), dove ogni mappa è calcolata in parallelo sull'intera immagine.

Lo stadio attentivo seleziona una locazione (x,y) e integra le features:

$$\text{Oggetto}(x,y) = \oplus_k F_k(x,y)$$

dove $\oplus$ rappresenta un'operazione di binding.

Senza attenzione alla locazione (x,y) , le features possono essere rilevate individualmente ma non correttamente associate, portando a "illusory conjunctions" (congiunzioni illusorie) dove features di oggetti diversi sono erroneamente combinate.

La probabilità di una congiunzione illusoria aumenta con il numero di oggetti N :

$$P(\text{illusory conjunction}) \propto N(N-1)/2$$

riflettendo il numero di possibili mis-bindings tra N oggetti.

Modello di Bundesen: Theory of Visual Attention (TVA)

La TVA di Claus Bundesen fornisce un framework matematico formale per l'attenzione. In questo modello, gli oggetti competono per essere categorizzati e entrare nella memoria visiva a breve termine.

La forza di evidenza $v(x,i)$ che l'oggetto x appartiene alla categoria i è:

$$v(x,i) = \eta(x,i) \beta_i w_x$$

dove:

- $\eta(x,i)$ è la forza sensoriale dell'evidenza (determinata bottom-up dalle features dell'oggetto)
- β_i è il peso decisionale per la categoria i (determinato top-down dal task)
- w_x è il peso attentivo per l'oggetto x (determinato da una combinazione di fattori bottom-up e top-down)

La probabilità che l'oggetto x sia categorizzato come i al tempo t è:

$$P(x \rightarrow i, t) = 1 - \exp(-v(x,i) \cdot t)$$

Il tasso totale di processing è:

$$C = \sum_x \sum_i v(x,i)$$

che rappresenta la capacità attentiva totale (misurata in elementi/secondo).

Per N oggetti competitivi, il tasso di processing per ciascuno è ridotto proporzionalmente:

$$v_{\text{efficace}}(x,i) = v(x,i) \cdot C / (C + \sum_{y \neq x} \sum_j v(y,j))$$

Questo modello predice accuratamente una vasta gamma di fenomeni attentivi, inclusi effetti di set size, attentional capture, e performance in compiti di ricerca visiva.

Quantificazione della capacità attentiva

Un approccio influente alla quantificazione dei limiti attentivi deriva dalla teoria del carico percettivo di Nilli Lavie. Questa teoria propone che la capacità di elaborazione percettiva sia fissa e limitata, e che questa capacità venga esaurita dal processing dei task rilevanti. Quando il carico percettivo è basso, la capacità residua viene automaticamente allocata al processing di stimoli irrilevanti (distractors). Quando il carico è alto, tutta la capacità è consumata dal task rilevante, eliminando il processing dei distractors.

La capacità può essere concettualizzata come una risorsa C che deve essere distribuita tra n stimoli:

$$\sum_{i=1}^n c_i \leq C$$

dove c_i è la capacità allocata allo stimolo i . Esperimenti comportamentali suggeriscono che C è nell'ordine di 40-60 Hz se misurato in termini di frequenza critica di fusione per stimoli multipli, o equivalentemente può processare circa 3-5 oggetti discreti simultaneamente con attenzione focalizzata.

Teoria della risorsa attentiva come informazione mutua

Un approccio information-theoretic alla capacità attentiva è stato proposto da diversi ricercatori. In questo framework, la capacità attentiva è quantificata come la massima informazione mutua tra stimoli e risposte.

Sia S l'insieme di possibili stimoli e R l'insieme di possibili risposte. L'informazione mutua è:

$$I(S;R) = \sum_s \sum_r P(s,r) \log_2(P(s,r)/(P(s)P(r)))$$

La capacità attentiva è:

$$C_{att} = \max_P(s) I(S;R)$$

Per un canale con rumore gaussiano additivo di varianza σ^2_n e segnale di potenza σ^2_s :

$$C_{att} = (1/2) \log_2(1 + \sigma^2_s/\sigma^2_n)$$

Questo è analogo alla formula di Shannon per la capacità del canale.

L'attenzione aumenta σ^2_s (amplifica il segnale) e/o riduce σ^2_n (riduce il rumore), aumentando la capacità del canale per stimoli attenzionati.

Misurazioni psicofisiche suggeriscono che l'attenzione può aumentare la capacità di circa 2-4 bit per stimolo attenzionato, corrispondente a un aumento del SNR di circa 4-16 volte.

Il paradigma dell'attentional blink

Un fenomeno che rivela i limiti temporali dell'attenzione è l'attentional blink. In questo paradigma, i partecipanti osservano una sequenza rapida di stimoli (tipicamente lettere o numeri) presentati alla stessa posizione spaziale, con una frequenza di circa 10 stimoli al secondo. Il compito richiede l'identificazione di due target (ad esempio, lettere in rosso in una sequenza di lettere nere).

Il risultato notevole è che quando il secondo target (T2) appare entro circa 200-500 ms dopo il primo target (T1), i partecipanti spesso falliscono nell'identificarlo, anche se T1 è stato correttamente identificato. La probabilità di riportare correttamente T2 in funzione dell'intervallo tra T1 e T2 (stimulus onset asynchrony, SOA) mostra un pattern caratteristico:

$$P(T2|T1) \approx 0,2-0,4 \quad \text{per SOA} = 100-300 \text{ ms}$$

$$P(T2|T1) \approx 0,9-0,95 \quad \text{per SOA} > 500 \text{ ms}$$

Questo pattern temporale suggerisce che l'elaborazione attentiva di T1 rende il sistema attentivo temporaneamente refrattario all'elaborazione di T2. Il periodo di refrattarietà può essere modellato come un periodo di consolidamento durante il quale T1 viene trasferito dalla memoria sensoriale alla memoria di lavoro.

Modello della bottleneck temporale

Il modello più influente dell'attentional blink propone un collo di bottiglia nell'accesso alla memoria di lavoro. Questo può essere formalizzato come segue:

Assumiamo che il consolidamento di un target in memoria di lavoro richieda un'operazione seriale che dura $\tau_{\text{cons}} \approx 500 \text{ ms}$. Durante questo periodo, nessun altro target può essere consolidato.

Sia t_1 il tempo di presentazione di T1 e t_2 quello di T2, con $\Delta t = t_2 - t_1$.

Se $\Delta t < \tau_{\text{cons}}$, T2 è presentato mentre il sistema sta ancora consolidando T1. In questo caso:

$$P(T2 \text{ riportato} | T1 \text{ riportato}) = P(T2 \text{ escapes blink})$$

Diversi modelli specificano $P(T2 \text{ escapes blink})$ diversamente:

Modello a soglia: T2 sfugge al blink solo se è presentato dopo che il consolidamento di T1 è completato:

$$P(T2 \text{ escapes}) = 1 \quad \text{se } \Delta t \geq \tau_{\text{cons}}$$

$$P(T2 \text{ escapes}) = 0 \quad \text{se } \Delta t < \tau_{\text{cons}}$$

Modello graduale: La probabilità aumenta gradualmente:

$$P(T2 \text{ escapes}) = 1 - \exp(-\Delta t / \tau_{\text{cons}})$$

Modello di interferenza: T2 può interrompere il consolidamento di T1, con probabilità che dipende dalla forza relativa:

$$P(T2 \text{ escapes}) = S_{T2} / (S_{T1} + S_{T2})$$

dove S_{T1} , S_{T2} sono le "forze" dei due target.

I dati empirici supportano meglio i modelli graduali o di interferenza rispetto al modello a soglia rigida.

Lag-1 sparing

Un'eccezione interessante al pattern dell'attentional blink è il fenomeno "lag-1 sparing": quando T2 appare immediatamente dopo T1 ($SOA \approx 100$ ms, $lag = 1$ item), la performance su T2 è spesso alta, evitando il blink.

Questo può essere spiegato assumendo che l'attenzione può rimanere "aperta" per un breve periodo dopo T1, permettendo a T2 di essere processato nello stesso "episodio attentivo". Matematicamente:

$$P(T2|T1, lag=1) \approx P(T1 \text{ e } T2 \text{ processati insieme})$$

Se T2 cade nella finestra temporale di integrazione (≈ 100 -150 ms), può essere consolidato insieme a T1 come un singolo episodio.

Modelli computazionali dell'allocazione attentiva

I modelli computazionali moderni dell'attenzione visiva, come il modello di saliency map di Laurent Itti e Christof Koch, propongono che l'attenzione sia guidata da una mappa di saliency che integra informazione bottom-up (basata su caratteristiche visive come contrasto, colore, orientamento) e top-down (basata su task goals e aspettative).

La saliency map $S(x,y)$ assegna un valore di saliency ad ogni posizione (x,y) nella scena visiva:

$$S(x,y) = \sum_i w_i \times F_i(x,y)$$

dove $F_i(x,y)$ rappresenta una feature map per la caratteristica i (ad esempio, contrasto di luminanza, contrasto di colore, presenza di bordi ad alto contrasto), e w_i è il peso di quella feature nella determinazione della saliency complessiva.

La probabilità che l'attenzione venga allocata a una particolare posizione è proporzionale alla saliency:

$$P(\text{attenzione a } (x,y)) = S(x,y) / \iint S(x',y') dx'dy'$$

Questi modelli prevedono con accuratezza moderata (circa 70-80% di concordanza) dove le persone dirigono la loro attenzione durante la visualizzazione libera di immagini statiche. Tuttavia, essi catturano principalmente l'aspetto bottom-up dell'attenzione e sono meno efficaci nel predire l'attenzione in task guidati da obiettivi specifici.

Implementazione computazionale del modello di Itti-Koch

Il modello di Itti e Koch calcola la saliency attraverso i seguenti passi:

1. **Estrazione di features multi-scala:** L'immagine è processata attraverso una piramide gaussiana a 9 scale. Per ogni scala $\sigma \in \{0, 1, \dots, 8\}$, si calcola:

$$I_{\sigma} = G_{\sigma} * I$$

dove G_{σ} è un filtro gaussiano con deviazione standard σ e $*$ indica convoluzione.

2. **Calcolo di center-surround differences:** Per diverse combinazioni di scale (c,s) dove $c \in \{2,3,4\}$ è il centro e $s = c + \delta$ con $\delta \in \{3,4\}$ è il surround:

$$\text{Contrasto_intensità}(c,s) = |I_c - I_{s\uparrow}|$$

$$\text{Contrasto_colore}(c,s) = |(R_c - G_c) - (R_s - G_s)\uparrow| + |(B_c - Y_c) - (B_s - Y_s)\uparrow|$$

$$\text{Contrasto_orientamento}_{\theta}(c,s) = |O_{\theta,c} - O_{\theta,s\uparrow}|$$

dove $\uparrow$ indica upsampling alla risoluzione più fine e O_{θ} sono mappe di orientamento per $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

3. **Normalizzazione e combinazione:** Ciascuna feature map è normalizzata a un range [0,1] e poi moltiplicata per $(1 - \max_locale/\max_globale)^2$ per sopprimere features uniformemente distribuite:

$$N(F) = (M - \bar{m})^2 \cdot F / M^2$$

dove M è il massimo globale e $\bar{m}$ è la media dei massimi locali.

4. **Integrazione cross-scale:** Le mappe a diverse scale sono combinate:

$$\bar{I} = \oplus_{\{c,s\}} N(\text{Contrasto_intensità}(c,s))$$

$$\bar{C} = \oplus_{\{c,s\}} N(\text{Contrasto_colore}(c,s))$$

$$\bar{O} = \sum_{\theta} \oplus_{\{c,s\}} N(\text{Contrasto_orientamento}_{\theta}(c,s))$$

dove $\oplus$ indica riduzione alla scala più fine e somma point-wise.

5. Mappa di saliency finale:

$$S = (\bar{I} + \bar{C} + \bar{O}) / 3$$

6. **Winner-take-all e inibizione di ritorno:** L'attenzione viene allocata iterativamente al massimo della saliency map, seguito da inibizione locale temporanea in quella regione.

Estensioni: Task-driven attention e apprendimento

Modelli più recenti estendono questo framework incorporando influenze top-down. Un approccio è weighted saliency:

$$S_{\text{task}}(x,y) = \sum_i w_i(\text{task}) \times F_i(x,y)$$

dove i pesi $w_i(\text{task})$ dipendono dal task corrente.

I pesi possono essere appresi da dati di eye-tracking usando approcci di machine learning. Per esempio, regressione lineare:

$$w^* = \text{argmin}_w \sum_n ||F_n w - y_n||^2 + \lambda ||w||^2$$

dove F_n sono le feature maps per l'immagine n , y_n è il pattern di fissazioni osservate, e λ è un termine di regolarizzazione.

Approcci più sofisticati usano deep learning, dove l'intera pipeline di estrazione features e pesatura è appresa end-to-end da dati di eye-tracking su migliaia di immagini.

La matematica delle rappresentazioni sparse

Teoria della codifica efficiente

Una prospettiva potente per comprendere perché il sistema visivo mantiene rappresentazioni sparse deriva dalla teoria della codifica efficiente, originariamente proposta da Horace Barlow negli anni 1960. Questa teoria suggerisce che i sistemi sensoriali si sono evoluti per rappresentare l'informazione sensoriale nel modo più efficiente possibile, minimizzando la ridondanza e massimizzando l'informazione trasmessa attraverso canali neurali limitati.

Dal punto di vista della teoria dell'informazione, la ridondanza R in una rappresentazione può essere quantificata come:

$$R = 1 - H(X) / H_{\max}(X)$$

dove $H(X)$ è l'entropia effettiva della rappresentazione X , e $H_{\max}(X)$ è l'entropia massima possibile (che si otterrebbe se tutti i possibili stati fossero equiprobabili). Una rappresentazione efficiente minimizza R , il che equivale a massimizzare $H(X)$ data la capacità del canale.

Per un set di N neuroni, ciascuno con F possibili firing rates, l'entropia massima sarebbe:

$$H_{\max} = N \times \log_2(F)$$

Tuttavia, se le firing rates dei neuroni sono correlate (come tendono ad essere nelle rappresentazioni ridondanti), l'entropia effettiva sarà inferiore. La codifica efficiente cerca di de-correlare le risposte neuronali, portando ad una rappresentazione in cui ogni neurone contribuisce informazione indipendente.

Quantificazione della ridondanza nelle immagini naturali

Le immagini naturali esibiscono correlazioni statistiche sostanziali. La funzione di autocorrelazione spaziale $C(\Delta x, \Delta y)$ per le immagini naturali è tipicamente:

$$C(\Delta x, \Delta y) = \sigma^2 \exp(-(\Delta x^2 + \Delta y^2)/(2\sigma^2_{\text{corr}}))$$

con $\sigma_{\text{corr}} \approx 2-5$ pixel per immagini ad alta risoluzione.

Lo spettro di potenza (trasformata di Fourier dell'autocorrelazione) segue approssimativamente una legge di potenza:

$$P(f) \propto 1/f^\alpha$$

con $\alpha \approx 2$ per immagini naturali. Questo implica che basse frequenze spaziali (strutture su larga scala) contengono la maggior parte dell'energia.

La ridondanza può essere quantificata confrontando l'entropia di un'immagine naturale con quella di un'immagine con pixel indipendenti:

$$R = 1 - H(\text{immagine_naturale}) / H(\text{pixel_indipendenti})$$

Per immagini a 8-bit dove $H(\text{pixel_indipendenti}) = 8 \text{ bit/pixel}$, misurazioni empiriche danno $H(\text{immagine_naturale}) \approx 1\text{-}2 \text{ bit/pixel}$ dopo accounting per correlazioni, quindi:

$$R \approx 1 - 1,5/8 \approx 0,81$$

Circa 81% di ridondanza, fornendo ampio spazio per compressione.

Sparse coding e rappresentazioni sovracomplete

Un approccio specifico alla codifica efficiente è lo sparse coding, in cui gli stimoli sensoriali sono rappresentati come combinazioni lineari di un dizionario di elementi di base (basis functions), con il vincolo che solo un piccolo numero di elementi del dizionario sia attivo per qualsiasi stimolo particolare.

Matematicamente, possiamo rappresentare uno stimolo visivo s come:

$$s \approx \sum_i a_i \times \varphi_i = \Phi a$$

dove φ_i sono gli elementi del dizionario (colonne della matrice Φ) e a_i sono i coefficienti di attivazione. Lo sparse coding impone che la maggior parte degli a_i siano zero o molto piccoli, con solo pochi coefficienti significativamente diversi da zero.

Il problema di trovare la rappresentazione sparse ottimale può essere formalizzato come un problema di ottimizzazione:

$$\text{minimizza } \|s - \Phi a\|^2 + \lambda \sum_i |a_i|$$

dove il primo termine misura l'errore di ricostruzione e il secondo termine (con λ come parametro di regolarizzazione) penalizza le rappresentazioni dense. La norma L1 ($\sum |a_i|$) favorisce soluzioni sparse rispetto alla norma L2.

Sorprendentemente, quando algoritmi di sparse coding vengono applicati a patch di immagini naturali, gli elementi del dizionario appreso assomigliano notevolmente ai campi recettivi dei neuroni nella corteccia visiva primaria V1. Questi campi recettivi sono tipicamente localizzati spazialmente, orientati, e selettivi per frequenze spaziali specifiche, caratteristiche corrispondenti ai filtri di Gabor.

Algoritmi per sparse coding

ISTA (Iterative Shrinkage-Thresholding Algorithm): Un metodo iterativo per risolvere il problema di sparse coding:

$$a^{(t+1)} = S_{\lambda}(a^{(t)} + \Phi^T(s - \Phi a^{(t)}))$$

dove S_λ è l'operatore di soft-thresholding:

$$S_\lambda(x) = \text{sign}(x) \max(|x| - \lambda, 0)$$

Questo iterativamente aggiorna i coefficienti, eseguendo gradient descent sull'errore di ricostruzione e poi applicando shrinkage per indurre sparsità.

Matching Pursuit: Un algoritmo greedy che seleziona iterativamente l'elemento del dizionario più correlato con il residuo corrente:

$$i^* = \arg\max_i |\varphi_i^T r(t)|$$

$$a_{i^*} = a_{i^*} + \varphi_{i^*}^T r(t)$$

$$r(t+1) = r(t) - (\varphi_{i^*}^T r(t)) \varphi_{i^*}$$

iniziando con residuo $r(0) = s$.

Learning del dizionario: Il dizionario Φ stesso può essere appreso dai dati. L'algoritmo K-SVD alterna tra:

1. **Sparse coding:** Fissa Φ , ottimizza a per tutti gli esempi di training
2. **Dictionary update:** Fissa a , aggiorna ciascuna colonna di Φ via SVD

Formalmente, per aggiornare φ_k :

$$E_k = S - \sum_{j \neq k} \varphi_j a_j^T$$

$$\varphi_k, a_k = \text{SVD}_1(E_k)$$

dove SVD_1 estrae la prima componente singolare.

Emergence di Gabor-like filters

I filtri di Gabor in 2D sono definiti come:

$$G(x, y; \theta, f, \sigma) = \exp(-(x'^2 + y'^2)/(2\sigma^2)) \cos(2\pi f x' + \psi)$$

dove:

- (x', y') sono coordinate ruotate di θ : $x' = x \cos \theta + y \sin \theta$, $y' = -x \sin \theta + y \cos \theta$
- f è la frequenza spaziale

- σ è la deviazione standard dell'involucro gaussiano
- γ è l'aspect ratio
- ψ è la fase

Quando sparse coding è applicato a patch 16×16 da immagini naturali, i filtri appresi mostrano:

- Localizzazione spaziale (supporto compatto)
- Selettività per orientamento (θ distribuito uniformemente)
- Selettività per frequenza (f tipicamente 0,1-0,5 cicli/pixel)
- Struttura Gabor-like con envelope gaussiano e carrier sinusoidale

La somiglianza con campi recettivi V1 è quantificabile. Misurazioni neurofisiologiche mostrano che neuroni V1 hanno:

- Dimensione campo recettivo: 0,5-2 gradi
- Selettività orientamento: bandwidth $\approx$ 30-40 gradi
- Selettività frequenza: bandwidth $\approx$ 1-1,5 ottave

I filtri appresi da sparse coding replicano queste statistiche con alta fedeltà, supportando l'ipotesi che V1 implementa una forma di sparse coding di immagini naturali.

Densità di attivazione nelle rappresentazioni neurali

Misurazioni neurofisiologiche supportano l'idea che le rappresentazioni corticali siano effettivamente sparse. Registrazioni da neuroni nella corteccia visiva durante la presentazione di immagini naturali mostrano che tipicamente solo il 5-20% dei neuroni in una popolazione è fortemente attivo in risposta a qualsiasi stimolo particolare.

Se definiamo la sparsità S di una rappresentazione come:

$$S = 1 - (\sum_i a_i / N)^2 / (\sum_i a_i^2 / N)$$

dove a_i sono le attivazioni dei N neuroni nella popolazione, otteniamo $S = 0$ per una rappresentazione completamente densa (tutti i neuroni ugualmente attivi) e $S \rightarrow 1$ per una rappresentazione completamente sparse (solo un neurone attivo).

Misurazioni empiriche in V1 danno valori di S tipicamente tra 0,7 e 0,9, indicando rappresentazioni altamente sparse. Questa sparsità aumenta progressivamente nelle aree visive superiori, con alcune aree temporali inferiori che mostrano $S > 0,95$.

Lifetime sparsity vs population sparsity

È importante distinguere due tipi di sparsità:

Population sparsity: Frazione di neuroni attivi in una popolazione per un dato stimolo. Se a_i è l'attivazione del neurone i per stimolo s :

$$S_{\text{pop}}(s) = (\# \text{ neuroni con } a_i > \theta) / N$$

Tipicamente $S_{\text{pop}} \approx 0,05-0,20$ in corteccia visiva.

Lifetime sparsity: Frazione di stimoli che attivano un dato neurone. Se $a_i(s)$ è l'attivazione del neurone i per stimolo s :

$$S_{\text{life}}(i) = (\# \text{ stimoli con } a_i(s) > \theta) / N_{\text{stimuli}}$$

Neuroni in IT (inferotemporal cortex) possono avere $S_{\text{life}} < 0,01$, rispondendo selettivamente a una categoria molto specifica di stimoli.

Benefici della sparsità

1. **Efficienza metabolica:** Il costo energetico è proporzionale al firing rate. Rappresentazioni sparse riducono firing rate medio.
2. **Capacità di memoria:** Memorie possono essere implementate come pattern di attivazione. Con population sparsity p , il numero di pattern ortogonali distinguibili è approssimativamente:

$$M \approx (N \text{ choose } pN) \approx (1/p)^{pN}$$

Per $N = 1000$, $p = 0,1$:

$$M \approx 10^{\{100\}}$$

vs. solo $\approx 2^{\{1000\}} \approx 10^{\{301\}}$ per rappresentazioni binarie dense (ma con overlap molto maggiore).

3. **Robustezza al rumore:** Con sparsità elevata, pattern diversi hanno overlap minimo, migliorando discriminabilità anche con rumore.
4. **Interpretabilità:** Neuroni sparsi tendono ad avere significato semantico più chiaro (es. "neurone di Jennifer Aniston").

Campionamento e ricostruzione della realtà

Il teorema del campionamento di Shannon-Nyquist

La teoria del campionamento fornisce vincoli matematici fondamentali su come l'informazione continua può essere rappresentata discretamente senza perdita di informazione. Il teorema del campionamento di Shannon-Nyquist afferma che un segnale continuo con frequenza massima $f_{\max}$ può essere perfettamente ricostruito da campioni discreti se il tasso di campionamento f_s soddisfa:

$$f_s \geq 2 \times f_{\max}$$

Questo limite di Nyquist stabilisce il tasso minimo di campionamento necessario per evitare aliasing. Nel contesto della visione, possiamo considerare l'arrangiamento spaziale dei fotorecettori come un processo di campionamento della distribuzione continua di luce nell'immagine retinica.

La densità di fotorecettori nella fovea è circa 200.000 coni/mm². Data la lunghezza focale dell'occhio di circa 17 mm, questo corrisponde a una risoluzione di circa 0,5 arcmin. Secondo il criterio di Nyquist, questo dovrebbe permettere di risolvere frequenze spaziali fino a circa 60 cicli/grado.

Tuttavia, l'acuità visiva misurata comportamentalmente raggiunge circa 30-40 cicli/grado in condizioni ottimali, suggerendo che fattori oltre il semplice campionamento retinico (come l'ottica dell'occhio e il processing neurale) limitano la risoluzione effettiva.

Aliasing e sottocampionamento nella periferia

Nella periferia retinica, la densità di fotorecettori è molto ridotta, violando il criterio di Nyquist per frequenze spaziali medie-alte. Questo porta ad aliasing: frequenze spaziali alte sono erroneamente rappresentate come frequenze più basse.

Per una densità di fotorecettori $p(e)$ all'eccentricità e , la frequenza di Nyquist è:

$$f_{\text{Nyquist}}(e) = p(e) / 2$$

Se la scena contiene componenti a frequenza $f > f_{\text{Nyquist}}(e)$, queste saranno alias a:

$$f_{\text{alias}} = |f - n \cdot 2f_{\text{Nyquist}}(e)|$$

per qualche intero n .

Questo spiega alcune illusioni percettive in periferia, come il "wagon wheel effect" dove ruote che ruotano rapidamente appaiono ruotare più lentamente o all'indietro.

Ricostruzione e interpolazione

Dato un segnale campionato $s[n] = s(nT)$ dove $T = 1/f_s$ è il periodo di campionamento, la ricostruzione del segnale continuo è ottenuta via:

$$s_{\text{ricostruito}}(t) = \sum_n s[n] \text{sinc}((t - nT)/T)$$

dove $\text{sinc}(x) = \sin(\pi x)/(\pi x)$ è il kernel di interpolazione ideale.

Nella pratica, il sistema visivo probabilmente usa interpolazione più semplice (es. lineare o cubica) piuttosto che sinc perfetto, introducendo distorsioni minori ma biologicamente plausibili.

Compressione e perdita di informazione

Mentre il campionamento retinico soddisfa approssimativamente il criterio di Nyquist nella fovea, i successivi stadi di elaborazione introducono compressione sostanziale. Il nervo ottico rappresenta un drammatico collo di bottiglia: i circa 126 milioni di fotorecettori convergono su circa 1 milione di cellule ganglionari retiniche.

Questo implica un rapporto di compressione di circa:

$$r_{\text{compressione}} = 126.000.000 / 1.000.000 = 126:1$$

Naturalmente, questa non è una compressione casuale. Le cellule ganglionari implementano trasformazioni specifiche dell'input dei fotorecettori, come la rilevazione di contrasti locali attraverso campi recettivi centro-surround, e la separazione di diversi canali di informazione (parvocellulare per dettagli ad alta risoluzione spaziale e colore, magnocellulare per movimento e cambiamenti temporali).

Possiamo stimare la quantità di informazione trasmessa dal nervo ottico. Se ogni cellula gangliare ha un tasso medio di firing di circa 40 spike/s e può trasmettere circa 2-3 bit di informazione per spike (basandosi su studi di teoria dell'informazione delle risposte neurali), la bandwidth totale sarebbe:

$$I_{\text{nervo_ottico}} = 1.000.000 \times 40 \times 2,5 = 100 \times 10^6 \text{ bit/s} = 12,5 \text{ MB/s}$$

Confrontato con l'input sensoriale grezzo stimato di 2,09 GB/s, questo rappresenta una riduzione di circa 167 volte. Gran parte di questa compressione è ottenuta rimuovendo correlazioni spaziali e temporali nell'input visivo, mantenendo principalmente informazione sui cambiamenti e sui contrasti.

Trasformazioni retinica: dal segnale alla feature

Le cellule ganglionari retiniche non semplicemente sottocampionano i fotorecettori ma eseguono trasformazioni elaborate. Il modello classico è il campo recettivo centro-surround:

$$R(x,y) = K_c \exp(-(x^2+y^2)/(2\sigma_c^2)) - K_s \exp(-(x^2+y^2)/(2\sigma_s^2))$$

dove K_c , K_s sono i pesi del centro e surround, e $\sigma_c < \sigma_s$.

La risposta della cellula gangliare a uno stimolo $I(x,y)$ è:

$$r = \iint R(x,y) I(x,y) dx dy$$

Questo è essenzialmente un filtro passa-banda che risponde fortemente a bordi e contrasti locali, sopprimendo illuminazione uniforme.

In termini di teoria dell'informazione, questa trasformazione rimuove correlazioni di basso ordine (redundancy reduction). L'informazione mutua tra input e output può essere calcolata:

$$I(\text{Input}; \text{Output}) = H(\text{Output}) - H(\text{Output}|\text{Input})$$

Per campi recettivi ottimizzati, $H(\text{Output})$ è vicino al massimo (decorrelazione), mentre $H(\text{Output}|\text{Input})$ è minimizzato (rumore basso).

Ricostruzione inferenziale

Data la natura frammentaria e compressa della rappresentazione interna, il cervello deve ricostruire una percezione coerente della scena attraverso processi inferenziali. Questo approccio è formalizzato nella teoria della percezione come inferenza bayesiana, proposta in varie forme da Hermann von Helmholtz, e più recentemente sviluppata da ricercatori come Knill e Richards.

Secondo questa prospettiva, la percezione P di uno stato del mondo W data l'evidenza sensoriale E può essere calcolata usando il teorema di Bayes:

$$P(W|E) = P(E|W) \times P(W) / P(E)$$

Il cervello combina la likelihood $P(E|W)$, che rappresenta quanto probabile è l'evidenza sensoriale dato uno particolare stato del mondo, con le prior $P(W)$, che rappresentano le aspettative precedenti su cosa è probabile nel mondo, per arrivare alla percezione posteriore $P(W|E)$.

Questa formulazione spiega numerosi fenomeni percettivi, compresa la resistenza al cambiamento. Se le prior sono forti (riflettendo una storia recente di stabilità della scena), allora anche evidenza sensoriale di un cambiamento può essere insufficiente a superare le prior e produrre una percezione del cambiamento.

Possiamo modellare questo quantitativamente. Se la scena è stata stabile per un tempo t , la prior per la stabilità diventa:

$$P(\text{stabile}) = 1 - \exp(-t/\tau_{\text{stabilità}})$$

dove $\tau_{\text{stabilità}}$ è una costante di tempo che riflette quanto rapidamente il sistema aggiorna le sue aspettative. Valori tipici di $\tau_{\text{stabilità}}$ basati su esperimenti comportamentali sono nell'ordine di 500-1000 ms per scene statiche.

Inferenza bayesiana ottimale con prior gerarchici

Un modello più sofisticato incorpora prior gerarchici su diversi livelli di astrazione:

Livello basso (features locali):

$$P(\text{feature_locale} \mid \text{immagine}) \propto P(\text{immagine} \mid \text{feature_locale}) P(\text{feature_locale})$$

Livello medio (oggetti):

$$P(\text{oggetto} \mid \text{features}) \propto P(\text{features} \mid \text{oggetto}) P(\text{oggetto})$$

Livello alto (scena globale):

$$P(\text{scena} \mid \text{oggetti}) \propto P(\text{oggetti} \mid \text{scena}) P(\text{scena})$$

L'inferenza può procedere bottom-up (dai pixel alle features agli oggetti alla scena) e top-down (aspettative sulla scena vincolano interpretazioni di oggetti e features).

Il posterior finale integra attraverso i livelli:

$$P(\text{scena, oggetti, features} \mid \text{immagine}) \propto P(\text{immagine} \mid \text{features}) P(\text{features} \mid \text{oggetti}) P(\text{oggetti} \mid \text{scena}) P(\text{scena})$$

Questa struttura gerarchica spiega come prior di alto livello ("scene di cucina tipicamente contengono fornelli") possono influenzare rilevazione di basso livello, rendendo più facile rilevare cambiamenti in oggetti expected ma più difficile rilevare cambiamenti in oggetti unexpected.

Predictive coding come implementazione neurale

Il framework di predictive coding propone un'implementazione neurale plausibile dell'inferenza bayesiana. In questo modello:

1. Neuroni in livelli alti della gerarchia mantengono ipotesi sul mondo (rappresentazioni generative)
2. Queste ipotesi generano predizioni che propagano verso il basso
3. Errori di predizione (differenza tra predizione e input effettivo) propagano verso l'alto
4. Il sistema minimizza errori di predizione aggiornando le rappresentazioni

Matematicamente, per una gerarchia a L livelli con stati latenti $x^{(1)}, \dots, x^{(L)}$:

$$x^{(l)} = f^{(l)}(x^{(l+1)}) + \varepsilon^{(l)}$$

dove $f^{(l)}$ è la funzione generativa e $\varepsilon^{(l)}$ è rumore di predizione.

L'errore di predizione al livello l è:

$$e^{(l)} = x^{(l)} - f^{(l)}(x^{(l+1)})$$

L'energia libera totale è:

$$F = \sum_l \left(\frac{1}{2} e^{(l)T} \Pi^{(l)} e^{(l)} \right)$$

dove $\Pi^{(l)}$ è la matrice di precisione (inversa della covarianza) per il livello l .

Il sistema minimizza F attraverso:

$$dx^{(l)}/dt = -\partial F / \partial x^{(l)} = \Pi^{(l)} e^{(l)} - (\partial f^{(l+1)} / \partial x^{(l)})^T \Pi^{(l+1)} e^{(l+1)}$$

Questo implementa un bilancio tra fitting dell'input corrente (primo termine) e consistenza con prior dal livello superiore (secondo termine).

Neuroscienze della change blindness

Correlati neurali della consapevolezza del cambiamento

Studi di neuroimaging hanno iniziato a identificare i correlati neurali della change blindness e della rilevazione del cambiamento. Esperimenti utilizzando fMRI (risonanza magnetica funzionale) durante compiti di change detection hanno rivelato diverse regioni cerebrali che mostrano attività differenziale quando i cambiamenti vengono rilevati versus quando passano inosservati.

Una rete frontoparietale, includendo la corteccia prefrontale dorsolaterale (dlPFC), la corteccia parietale posteriore (PPC), e la frontal eye field (FEF), mostra aumentata attività BOLD (Blood Oxygen Level Dependent) quando i cambiamenti sono rilevati

consapevolmente. Questa rete è consistente con le reti attentive identificate in altri paradigmi.

Importantemente, l'attività in regioni visive precoci come V1 e V2 non differisce significativamente tra cambiamenti rilevati e non rilevati, suggerendo che l'informazione sul cambiamento è disponibile nelle rappresentazioni visive iniziali ma non accede alla consapevolezza senza il coinvolgimento delle regioni frontoparietal.

Misurazioni quantitative mostrano che il segnale BOLD nella dIPFC aumenta di circa 0,3-0,5% quando i cambiamenti sono rilevati rispetto a quando non lo sono, e questo aumento è correlato con i report soggettivi di consapevolezza del cambiamento.

Analisi quantitativa di connettività funzionale

Studi di connettività funzionale usando fMRI hanno rivelato che la rilevazione consapevole del cambiamento è associata con aumentata sincronizzazione tra regioni frontoparietal e visive.

La connettività funzionale può essere quantificata attraverso correlazione delle serie temporali BOLD:

$$FC(A,B) = \text{corr}(\text{BOLD_A}(t), \text{BOLD_B}(t))$$

Durante rilevazione consapevole:

- $FC(\text{dIPFC}, \text{V4})$ aumenta da $\approx 0,2$ (baseline) a $\approx 0,45$
- $FC(\text{PPC}, \text{IT})$ aumenta da $\approx 0,15$ a $\approx 0,40$

Analisi di causalità Granger suggerisce che la direzione di influenza è primariamente frontoparietal $\rightarrow$ visiva durante rilevazione consapevole, consistente con feedback attentivo che modula processing visivo.

La causalità Granger da X a Y è definita come:

$$GC(X \rightarrow Y) = \log(\text{var}(\varepsilon_Y) / \text{var}(\varepsilon_{\{Y|X\}}))$$

dove ε_Y è l'errore di predizione di Y basato solo sulla sua storia passata, e $\varepsilon_{\{Y|X\}}$ include anche la storia di X.

Durante change detection riuscita:

- $GC(\text{dIPFC} \rightarrow \text{V4}) \approx 0,08$ (significativo, $p < 0,001$)
- $GC(\text{V4} \rightarrow \text{dIPFC}) \approx 0,02$ (non significativo)

Dinamiche temporali: evidenza da EEG e MEG

Gli studi elettroencefalografici (EEG) e magnetoencefalografici (MEG) forniscono una risoluzione temporale superiore per tracciare le dinamiche dell'elaborazione del cambiamento. Questi studi hanno identificato diversi componenti evento-correlati (ERP) associati con la rilevazione del cambiamento.

La componente **P300**, un'onda positiva che raggiunge il picco circa 300-400 ms dopo uno stimolo rilevante per il task, è consistentemente osservata quando i cambiamenti sono rilevati consapevolmente ma è assente o significativamente ridotta quando i cambiamenti non sono rilevati. L'ampiezza della P300 è tipicamente 5-10 μV per cambiamenti rilevati versus 1-2 μV per cambiamenti non rilevati, misurata a elettrodi parietali centrali.

Una componente più precoce, la **visual mismatch negativity (vMMN)**, appare circa 100-200 ms dopo il cambiamento e sembra riflettere il processing automatico della devianza visiva. Crucialmente, la vMMN è osservata anche per cambiamenti che non entrano nella consapevolezza, suggerendo che il sistema visivo registra il cambiamento a livello pre-attentivo, ma questo segnale non è sufficiente per la consapevolezza senza l'ulteriore elaborazione riflessa nella P300.

Analisi tempo-frequenza delle oscillazioni neurali

Oltre ai componenti ERP, l'analisi tempo-frequenza rivela cambiamenti nelle oscillazioni neurali durante change detection.

La potenza oscillatoria in diverse bande può essere calcolata usando trasformata wavelet:

$$W(t,f) = \int s(\tau) \psi^*_{\{t,f\}}(\tau) d\tau$$

dove $\psi_{\{t,f\}}$ è una wavelet madre (es. Morlet) centrata al tempo t e frequenza f .

Risultati empirici:

1. **Banda theta (4-8 Hz):** Aumento di potenza in corteccia frontale durante rilevazione consapevole
 - Power increase: +20-40% rispetto a baseline
 - Latenza: 200-400 ms post-cambiamento
2. **Banda alpha (8-12 Hz):** Desincronizzazione in corteccia visiva occipitale
 - Power decrease: -15-25%
 - Interpretazione: rilascio di inibizione, aumento di processing
3. **Banda gamma (30-80 Hz):** Aumento in rete frontoparietale
 - Power increase: +30-60%
 - Latenza: 250-450 ms
 - Interpretazione: binding di features in rappresentazione coerente

Sincronizzazione cross-frequenza: Phase-amplitude coupling tra theta e gamma:

$$PAC = |\langle A_{\gamma}(t) \exp(i\phi_{\theta}(t)) \rangle|$$

dove A_{γ} è l'ampiezza della gamma e ϕ_{θ} è la fase della theta.

Durante rilevazione consapevole, PAC aumenta significativamente, suggerendo che oscillazioni theta frontali modulano bursts di gamma in regioni visive, possibilmente coordinando processing distribuito.

Modulazione attentiva delle risposte neurali

Registrazioni single-unit da corteccia visiva di primati non umani hanno dimostrato che l'attenzione modula le risposte neurali in modo quantificabile. Quando l'attenzione è diretta a uno stimolo nel campo recettivo di un neurone, il tasso di firing del neurone può aumentare di un fattore di 1,3-2,5 (aumento del 30-150%) rispetto a quando l'attenzione è diretta altrove.

Questa modulazione può essere modellata come una moltiplicazione del gain:

$$r_{\text{atteso}} = g_{\text{att}} \times r_{\text{baseline}}$$

dove r_{atteso} è il tasso di firing atteso, r_{baseline} è il tasso senza attenzione, e g_{att} è il gain attentivo (tipicamente $g_{\text{att}} \in [1,3, 2,5]$).

Oltre all'aumento del gain, l'attenzione riduce anche la variabilità delle risposte neurali, migliorando effettivamente il rapporto segnale-rumore. La riduzione della varianza è tipicamente nell'ordine del 10-30%.

Questi effetti di modulazione attentiva forniscono un meccanismo neurale attraverso cui gli stimoli attesi ricevono elaborazione preferenziale, mentre gli stimoli non attesi (come i cambiamenti in regioni non attese) ricevono elaborazione degradata.

Modelli di modulazione attentiva: gain vs tuning

Due principali modelli spiegano come l'attenzione modula risposte neurali:

Modello di gain moltiplicativo: L'attenzione moltiplica la curva di tuning del neurone per un fattore costante:

$$r_{\text{att}}(s) = g \times r(s)$$

dove s è lo stimolo e $r(s)$ è la curva di tuning base.

Modello di sharpening del tuning: L'attenzione restringe la curva di tuning, aumentando selettività:

$$r_{att}(s) = r_{max} \exp(-((s - s_{pref})^2)/(2\sigma_{att}^2))$$

con $\sigma_{att} < \sigma_{baseline}$.

Evidenza empirica: Dati da V4 e MT suggeriscono un modello ibrido dove entrambi i meccanismi operano, con predominanza di gain in alcune popolazioni e sharpening in altre.

Un modello unificato:

$$r_{att}(s) = g \times r_{baseline}(s; \sigma_{att})$$

dove sia g che σ sono modulati dall'attenzione.

Information-theoretic analysis della modulazione attentiva

La modulazione attentiva può essere quantificata in termini di informazione mutua tra stimolo e risposta:

$$I(S;R) = H(R) - H(R|S)$$

L'attenzione aumenta $I(S;R)$ attraverso due meccanismi:

1. **Riduzione del rumore:** Diminuisce $H(R|S)$ (variabilità condizionale)
2. **Aumento della dinamica:** Aumenta $H(R)$ (range di risposte)

Misurazioni in V4 mostrano che l'attenzione aumenta $I(S;R)$ di circa 10-30%, corrispondente a un aumento di circa 0,3-0,8 bit per spike.

Implicazioni per la rappresentazione della realtà

La natura costruttiva della percezione

L'evidenza dalla change blindness e dai limiti attentivi supporta fortemente una visione costruttivista della percezione, in cui l'esperienza cosciente non è un'immagine fedele della realtà esterna ma una costruzione attiva basata su campionamento selettivo e inferenza.

Possiamo quantificare il grado di costruzione necessario. Se assumiamo che la rappresentazione interna mantiene dettagli ad alta risoluzione solo per circa il 5% della scena visiva in ogni momento (basandosi sui limiti di capacità della VSTM e sulla dimensione della regione foveale), e che i restanti 95% sono rappresentati solo schematicamente, allora per una scena tipica contenente circa 10^4 - 10^5 elementi distinguibili, il sistema mantiene rappresentazioni dettagliate di solo:

$$N_{dettagliato} = 0,05 \times 10^4 = 500 \text{ elementi}$$

Il resto deve essere riempito attraverso inferenza basata su conoscenza precedente, regolarità statistiche delle scene naturali, e informazione contestuale.

Filling-in e completion percettiva

Il sistema visivo dimostra capacità di filling-in: completare regioni dove informazione diretta è assente. Esempi includono:

1. **Blind spot:** Il punto cieco nella retina (dove il nervo ottico esce) non produce "buco" percettivo
2. **Bordi illusori:** Triangolo di Kanizsa - bordi percepiti dove non ci sono gradienti di luminanza
3. **Texture completion:** Pattern regolari vengono completati attraverso occlusioni

Questi possono essere modellati come problemi di interpolazione. Data informazione in regioni A e B, ma non in regione C tra loro, il sistema inferisce:

$$I_C = \operatorname{argmin}_I [\int_C \|\nabla I\|^2 dx dy + \lambda \int_{\partial C} (I - I_{\text{observed}})^2 ds]$$

Il primo termine favorisce smoothness (minimizza gradienti), il secondo assicura consistenza con boundary observations.

Questo è essenzialmente un problema di Poisson:

$$\nabla^2 I = 0 \text{ in } C$$

$$I = I_{\text{observed}} \text{ su } \partial C$$

Neuralmente, questo potrebbe essere implementato attraverso connessioni laterali che diffondono attivazione da regioni con input a regioni senza.

Statistiche delle scene naturali

La capacità del sistema visivo di mantenere rappresentazioni sparse funziona in parte perché le scene naturali hanno strutture statistiche regolari che permettono predizione e compressione efficiente. Le immagini naturali mostrano caratteristiche statistiche distintive:

1. **Distribuzione di scala 1/f delle frequenze spaziali:** Lo spettro di potenza delle immagini naturali segue approssimativamente una legge di potenza:

$$P(f) \propto 1/f^\alpha$$

dove f è la frequenza spaziale e α è tipicamente vicino a 2 per immagini naturali. Questo significa che le basse frequenze (strutture su larga scala) dominano l'energia dell'immagine.

2. **Correlazioni locali:** I pixel vicini tendono ad avere valori simili. La funzione di autocorrelazione delle immagini naturali decresce approssimativamente esponenzialmente con la distanza:

$$C(d) = C_0 \times \exp(-d/d_0)$$

dove d_0 è tipicamente circa 2-5 pixel per immagini ad alta risoluzione.

3. **Struttura sparsa dei bordi:** Mentre i bordi (cambiamenti bruschi di intensità o colore) contengono la maggior parte dell'informazione strutturale in un'immagine, essi costituiscono solo una piccola frazione dei pixel totali, tipicamente $< 5\%$.

Queste regolarità statistiche significano che le immagini naturali sono altamente comprimibili. Algoritmi di compressione standard come JPEG possono raggiungere rapporti di compressione di 10:1 a 20:1 con perdita percettiva minima, sfruttando precisamente queste regolarità.

Il sistema visivo, avendo evoluto in un mondo di immagini naturali, può aver internalizzato queste statistiche come prior bayesiane che guidano l'inferenza percettiva. Questo permetterebbe al sistema di ricostruire dettagli non direttamente campionati basandosi su regolarità apprese.

Prior statistici appresi dalle scene naturali

Il cervello potrebbe mantenere prior impliciti sulle statistiche di scene naturali, appresi attraverso esperienza. Questi prior possono essere caratterizzati empiricamente analizzando database di immagini naturali.

Distribuzione di orientamenti: In scene naturali, bordi verticali e orizzontali sono più comuni di obliqui:

$$P(\theta) \propto \exp(-\kappa|\sin(2\theta)|)$$

dove $\kappa > 0$ favorisce cardinali. Questo è consistente con l'"oblique effect" - maggiore acuità per orientamenti cardinali.

Co-occorrenza di features: Certe combinazioni di features co-occorrono frequentemente. Per esempio, "cielo blu" e "erba verde" spesso appaiono insieme. Questo può essere catturato da prior congiunti:

$$P(\text{feature_1}, \text{feature_2}) \neq P(\text{feature_1})P(\text{feature_2})$$

Apprendere questi prior permette inference più accurata quando informazione è ambigua.

Distribuzione di profondità: Scene naturali mostrano bias verso certe strutture 3D (es. superfici planari, simmetria, convessità). Prior su queste possono risolvere ambiguità nell'interpretazione di cues di profondità.

Costi e benefici delle rappresentazioni sparse

La strategia di mantenere rappresentazioni sparse ha sia costi che benefici che possono essere analizzati quantitativamente:

Benefici energetici: Il cervello consuma circa 20% dell'energia metabolica totale del corpo, nonostante rappresenti solo circa 2% della massa corporea. La corteccia visiva costituisce una frazione sostanziale del cervello (circa 15-20% della corteccia cerebrale totale). Rappresentazioni sparse riducono il costo energetico dell'elaborazione neurale.

Il costo energetico E del firing neuronale può essere stimato come:

$$E = N \times r \times e_{\text{spike}}$$

dove N è il numero di neuroni, r è il tasso medio di firing, e e_{spike} è l'energia per spike (circa 10^9 molecole di ATP per spike). Per una popolazione di 10^9 neuroni visivi, una riduzione del firing rate medio dal 20% al 5% (dovuta a rappresentazioni sparse) risparmierebbe:

$$\Delta E = 10^9 \times (0,20 - 0,05) \times 40 \text{ Hz} \times 10^9 \text{ ATP/s} = 6 \times 10^{18} \text{ ATP/s}$$

Questo rappresenta un risparmio energetico sostanziale.

Costi di accuratezza: Il costo principale delle rappresentazioni sparse è la perdita di informazione sui dettagli non rappresentati. Questo si manifesta nella change blindness: cambiamenti in regioni non attese o non attenzionate passano inosservati perché non sono rappresentati con sufficiente dettaglio per permettere il confronto.

Possiamo quantificare questo costo come la probabilità di mancare un cambiamento rilevante:

$$P(\text{miss}) = (1 - f_{\text{att}}) \times P(\text{cambiamento fuori dal focus})$$

dove f_{att} è la frazione della scena coperta dall'attenzione focalizzata (tipicamente 0,01-0,05). Per una scena con cambiamenti distribuiti uniformemente, questo darebbe:

$$P(\text{miss}) \approx 0,95-0,99$$

consistente con i risultati sperimentali della change blindness.

Analisi ottimalità evolutiva

Possiamo formalizzare il trade-off come problema di ottimizzazione evolutiva. Sia:

- C_{energy} : costo energetico per unità di densità rappresentazionale
- C_{miss} : costo di mancare informazione critica
- ρ : densità di rappresentazione (frazione di scene dettagliatamente rappresentata)

Il costo totale è:

$$C_{\text{total}}(\rho) = C_{\text{energy}} \times \rho + C_{\text{miss}} \times P_{\text{miss}}(\rho)$$

dove $P_{\text{miss}}(\rho)$ decresce con ρ (più densa la rappresentazione, meno errori).

Assumendo $P_{\text{miss}}(\rho) \approx \exp(-\kappa\rho)$:

$$C_{\text{total}}(\rho) = C_{\text{energy}} \times \rho + C_{\text{miss}} \times \exp(-\kappa\rho)$$

L'ottimo evolutivo ρ^* minimizza C_{total} :

$$dC_{\text{total}}/d\rho = C_{\text{energy}} - \kappa C_{\text{miss}} \exp(-\kappa\rho) = 0$$

Risolvendo:

$$\rho^* = (1/\kappa) \log(\kappa C_{\text{miss}} / C_{\text{energy}})$$

Per parametri realistici ($C_{\text{miss}} \approx 10 \times C_{\text{energy}}$, $\kappa \approx 20$):

$$\rho^* \approx (1/20) \log(200) \approx 0,26$$

Circa 26% di densità ottimale, non lontano dalle stime empiriche di $\approx 5\text{-}20\%$, suggerendo che sparsità osservata potrebbe essere near-optimal data i vincoli evolutivi.

Change blindness nel mondo reale: implicazioni pratiche

Sicurezza e vigilanza

La change blindness ha implicazioni critiche per compiti di sicurezza dove operatori umani devono rilevare cambiamenti o anomalie in ambienti complessi.

Airport security screening: Studi hanno mostrato che screeners mancano circa 30-50% di oggetti minacciosi in bagagli simulati, particolarmente quando il carico di lavoro è alto.

Il tasso di miss può essere modellato in funzione di:

- Numero di items nel bagaglio (N)
- Tempo di ispezione disponibile (T)
- Prevalenza di minacce (rarietà aumenta miss rate - "prevalence effect")

$$P(\text{miss}) = 1 - (1 - e^{-\lambda N}) (T/(T + \tau)) (1 + \beta(1 - \text{prev}))$$

dove λ , τ , β sono parametri empirici.

Strategie di mitigazione:

1. **Aumento tempo ispezione:** $T \uparrow$ riduce miss rate, ma con rendimenti decrescenti
2. **Dual-check systems:** Due operatori indipendenti riducono $P(\text{miss})^2 \approx 0,09$ se $P(\text{miss}) = 0,3$
3. **AI-assisted detection:** Sistemi automatici pre-screening riducono N effettivo

Interfacce utente e design

Change blindness in UI updates: Utenti spesso non notano cambiamenti in interfacce web/app, anche quando importanti.

Principi di design:

1. **Motion cues:** Animazioni di transizione catturano attenzione (aumenta rilevazione ~2-3×)
2. **Spatial consistency:** Non spostare elementi frequentemente
3. **Explicit notification:** Per cambiamenti critici, notifiche esplicite necessarie

Quantificazione: A/B testing può misurare change blindness:

$$\text{Conversion_rate} = f(\text{noticeability})$$

$$\text{Noticeability} \propto \text{salience} \times \text{duration} \times \text{centrality}$$

Ottimizzare questi fattori aumenta engagement.

Testimonianza oculare

La change blindness solleva questioni sull'affidabilità della testimonianza oculare. Testimoni possono essere estremamente fiduciosi anche quando i loro ricordi sono incompleti o errati.

Modello di confidence-accuracy:

Accuracy = $f(\text{encoding_quality}, \text{retention_interval}, \text{retrieval_context})$

Confidence = $g(\text{subjective_experience}, \text{social_pressure})$

Problematicamente, Accuracy e Confidence sono spesso non correlate ($r \approx 0,2-0,3$), specialmente per dettagli periferici.

Weapon focus effect: Durante crimini con armi, attenzione è catturata dall'arma, riducendo encoding di altri dettagli (es. volto perpetratore).

Modellabile come:

Attention(weapon) $\uparrow \rightarrow$ Attention(face) $\downarrow \rightarrow$ Encoding(face) $\downarrow \rightarrow$ Accuracy(face) $\downarrow$

Empiricamente, identificazione da lineup è $\approx 50-60\%$ accurata con arma presente vs $\approx 70-80\%$ senza.

Meccanismi compensatori e strategie di superamento

Strategie attentive seriali

Sebbene il sistema attentivo sia limitato nella quantità di informazione che può processare in parallelo, può compensare parzialmente attraverso strategie seriali in cui l'attenzione viene spostata sequenzialmente tra diverse regioni o aspetti della scena.

La velocità di questi shift attentivi è limitata. Gli shift attentivi espliciti, accompagnati da movimenti oculari saccadici, possono avvenire circa 3-5 volte al secondo. Gli shift attentivi coperti (senza movimenti oculari) possono essere più rapidi, fino a circa 10 shift/s in condizioni ottimali.

Se modelliamo il campionamento attentivo come un processo seriale con tasso r_{shift} , il numero totale di regioni che possono essere campionate in un tempo T è:

$$N_{\text{campionato}} = r_{\text{shift}} \times T$$

Per una scena contenente N_{totale} regioni potenzialmente rilevanti, la frazione campionata sarebbe:

$$f_{\text{campionato}} = (r_{\text{shift}} \times T) / N_{\text{totale}}$$

Per un compito di change detection con $T = 10$ s e $r_{\text{shift}} = 4$ Hz, se la scena contiene 100 regioni potenzialmente rilevanti:

$$f_{\text{campionato}} = (4 \times 10) / 100 = 0,4$$

Questo significa che anche con campionamento seriale attivo, solo circa il 40% delle regioni può essere campionato in 10 secondi, lasciando il 60% non ispezionato e quindi suscettibile a change blindness.

Strategie di ricerca ottimali

Data limitata capacità di campionamento, quali regioni dovrebbero essere priorizzate? Questo può essere formulato come problema di decision-making ottimale.

Sia V_i il valore atteso di ispezionare la regione i (probabilità che contenga cambiamento $\times$ importanza del cambiamento). Il problema è selezionare sequenza di regioni per massimizzare valore totale atteso sotto vincolo temporale.

Greedy strategy: Seleziona sempre la regione con V_i più alto non ancora ispezionata:

$$i^* = \operatorname{argmax}_i V_i$$

Optimal (intratractable): Risolvere il problema completo di programmazione dinamica considerando tutte le sequenze possibili.

Empiricamente, umani approssimano strategie greedy ma con bias verso:

- Regioni centrali (bias di centralità)
- Regioni ad alto contrasto (cattura bottom-up)
- Regioni semanticamente rilevanti (guidance top-down)

Utilizzo di conoscenza semantica

Il sistema visivo può parzialmente superare i limiti di capacità utilizzando conoscenza semantica per guidare l'attenzione verso regioni probabilmente rilevanti o probabilmente soggette a cambiamento.

Questo approccio può essere formalizzato come un campionamento guidato dalla saliency semantica:

$$P(\text{campionamento} | \text{regione}_i) \propto S_{\text{semantica}}(\text{regione}_i)$$

dove $S_{\text{semantica}}$ riflette l'importanza o la probabilità di cambiamento della regione basata su conoscenza precedente.

Esperimenti hanno dimostrato che la rilevazione del cambiamento è significativamente più rapida per cambiamenti in oggetti semanticamente centrali (quelli rilevanti per il significato o il "gist" della scena). Ad esempio, in un'immagine di una cucina, cambiamenti al fornello o al frigorifero vengono rilevati circa 2-3 volte più rapidamente rispetto a cambiamenti a oggetti periferici come decorazioni murali.

Apprendimento di prior semantici

I prior semantici possono essere appresi da esposizione statistica a scene. Un modello computazionale:

1. **Database di scene:** Corpus di N immagini annotate con oggetti e loro posizioni

Apprendimento di co-occorrenza: Per coppia di oggetti (A,B):

$$P(B|A) = \text{freq}(A,B) / \text{freq}(A)$$

2.

3. **Uso per guidance:** Quando A è identificato, aumenta prior per B in posizioni tipiche

Esempio: In scene di cucina:

- $P(\text{frigorifero} | \text{tavolo}) = 0,8$ (alta co-occorrenza)
- $P(\text{pneumatico} | \text{tavolo}) = 0,01$ (rara co-occorrenza)

Quindi, cambiamenti a frigoriferi sono più facilmente rilevati in cucine per guidance semantica.

Strategie di confronto esplicito

In compiti dove la rilevazione del cambiamento è critica, gli osservatori possono adottare strategie di confronto esplicito, in cui deliberatamente tentano di codificare caratteristiche specifiche della scena originale e poi confrontarle con la scena modificata.

Questa strategia aumenta sostanzialmente la performance nella change detection ma ha costi significativi:

1. **Carico cognitivo:** Richiede impegno attentivo e di memoria di lavoro sostenuti, riducendo le risorse disponibili per altri compiti.
2. **Tempo:** Il confronto esplicito è lento, con tempi di scanning che aumentano linearmente con il numero di regioni da confrontare:

$$T_{\text{confronto}} = t_0 + k \times N_{\text{regioni}}$$

dove t_0 è un tempo di base e k è il tempo per regione (tipicamente 200-500 ms/regione).

3. **Errori:** Anche con confronto esplicito, gli errori rimangono sostanziali, particolarmente quando le caratteristiche da confrontare sono numerose o complesse.

Training e expertise

Il training può migliorare performance di change detection sviluppando:

1. **Chunking:** Raggruppare features in unità maggiori riduce carico di memoria
2. **Schema knowledge:** Expert hanno schemi più ricchi che facilitano encoding
3. **Strategie sistematiche:** Pattern di scanning più efficienti

Evidenza da domini di expertise (es. radiologi, controllori di volo):

- Accuracy: +15-30% rispetto a novizi
- Tempo di detection: -20-40%
- Meccanismo: Non capacità attentiva maggiore, ma strategie più efficienti

Prospettive future: modelli computazionali integrati

Deep learning e attenzione

Recenti sviluppi in deep learning hanno prodotto modelli di visione artificiale che incorporano meccanismi di attenzione ispirati ai sistemi biologici. Le reti neurali con meccanismi di attenzione, come le Transformer utilizzate in modelli visione-linguaggio, implementano forme di selezione attentiva che presentano interessanti parallelismi con l'attenzione biologica.

In questi modelli, l'attenzione è implementata attraverso pesi dinamici che determinano quanto ogni elemento di input influenza l'output:

$$\text{Output} = \sum_i \alpha_i \times V_i$$

dove α_i sono i pesi attentivi (che sommano a 1) e V_i sono i valori delle diverse posizioni o feature. I pesi attentivi sono calcolati attraverso:

$$\alpha_i = \text{softmax}(Q \cdot K_i / \sqrt{d_k})$$

dove Q è una query, K_i sono chiavi associate a diverse posizioni, e d_k è la dimensionalità dello spazio delle chiavi.

Sorprendentemente, questi modelli mostrano fenomeni simili alla change blindness: quando addestrati su compiti di riconoscimento di immagini, spesso falliscono nel rilevare cambiamenti in regioni dell'immagine che non ricevono peso attentivo elevato, analogamente ai sistemi biologici.

Vision Transformers e attention maps

I **Vision Transformers (ViT)** applicano l'architettura Transformer a patches di immagini. Un'immagine è divisa in P patches, ciascuna embedded in un vettore:

$$x_i = \text{Linear}(\text{Patch}_i) + \text{Pos_emb}_i$$

Il self-attention calcola per ogni patch:

$$\text{Attention}(X) = \text{softmax}(XW_Q (XW_K)^T / \sqrt{d}) XW_V$$

Le attention maps risultanti rivelano quali patches il modello "attende" quando processa ciascuna patch.

Analisi di change blindness in ViT:

- Cambiamenti in patches con alta attention ($\alpha > 0,1$): rilevati $\approx 85\%$ dei casi
- Cambiamenti in patches con bassa attention ($\alpha < 0,01$): rilevati $\approx 25\%$ dei casi

Questo pattern è notevolmente simile ai dati umani, suggerendo che meccanismi di attenzione nelle reti neurali artificiali catturano aspetti dell'attenzione biologica.

Predizione e coding predittivo

Un framework teorico influente per comprendere la percezione è il predictive coding, che propone che il cervello mantenga un modello generativo del mondo e usi questo modello per predire l'input sensoriale. La percezione emerge dal confronto tra predizioni e input effettivo, con solo gli "errori di predizione" (discrepanze tra predetto e osservato) che vengono propagati verso l'alto nella gerarchia corticale.

In questo framework, un cambiamento nella scena verrà rilevato solo se genera un errore di predizione sufficientemente grande:

$$|\text{input_effettivo} - \text{predizione}| > \theta$$

dove θ è una soglia di rilevazione. Se la scena è stabile per un tempo esteso, il modello predittivo diventa sempre più accurato, e le predizioni si avvicinano all'input effettivo, rendendo difficile rilevare cambiamenti gradualmente.

La change blindness può essere interpretata come un fallimento nel generare errori di predizione sufficientemente grandi, particolarmente quando i cambiamenti avvengono in regioni che non sono attivamente predette (perché fuori dal focus attentivo) o quando avvengono durante transizioni (mascherati dalla discontinuità visiva).

Implementazione computazionale di predictive coding

Un modello implementabile di predictive coding per change detection:

Architettura:

- Livelli generativi: $g^l(l)$ che predicono attività nel livello inferiore
- Livelli di errore: $e^l(l) = x^l(l) - \hat{g}^l(l)(x^l(l+1))$

Dinamiche:

$$dx^l(l)/dt = -\lambda_x x^l(l) + f(x^l(l+1)) + \varepsilon^l(l)$$

$$d\varepsilon^l(l)/dt = -\lambda_\varepsilon \varepsilon^l(l) + (x^l(l) - f(x^l(l+1)))$$

Rilevazione cambiamento: Un cambiamento è rilevato quando:

$$||\varepsilon^l(l)|| > \theta_l \text{ per qualche livello } l$$

Risultati simulazioni:

- Cambiamenti in regioni predette (basso ε baseline): alta rilevazione
- Cambiamenti in regioni non predette (alto ε baseline): bassa rilevazione (mascherati da rumore)
- Cambiamenti durante transizioni: non rilevati (errors transienti trattati come rumore)

Questi pattern replicano change blindness umana.

Integrazione multimodale

Un'area di ricerca promettente riguarda come l'integrazione di informazione da multiple modalità sensoriali possa mitigare i limiti della percezione unimodale. Ad esempio, suoni associati con cambiamenti visivi possono attirare l'attenzione e facilitare la rilevazione del cambiamento.

Studi hanno dimostrato che quando un cambiamento visivo è accompagnato da un suono congruente, i tempi di rilevazione si riducono di un fattore di circa 2-3 rispetto alla condizione solo visiva. Questo suggerisce che segnali cross-modalità possono guidare efficacemente l'allocazione attentiva.

La matematica dell'integrazione multisensoriale può essere descritta attraverso modelli di combinazione ottimale, dove l'evidenza da diverse modalità viene pesata in base alla loro affidabilità relativa:

$$s_{\text{integrato}} = w_{\text{vis}} \times s_{\text{vis}} + w_{\text{aud}} \times s_{\text{aud}}$$

dove i pesi sono determinati dalle varianze delle stime:

$$w_{\text{vis}} = (1/\sigma^2_{\text{vis}}) / (1/\sigma^2_{\text{vis}} + 1/\sigma^2_{\text{aud}})$$

$$w_{\text{aud}} = (1/\sigma^2_{\text{aud}}) / (1/\sigma^2_{\text{vis}} + 1/\sigma^2_{\text{aud}})$$

Questo schema di integrazione ottimale minimizza la varianza della stima combinata:

$$\sigma^2_{\text{integrato}} = 1 / (1/\sigma^2_{\text{vis}} + 1/\sigma^2_{\text{aud}})$$

migliorando l'affidabilità della rilevazione del cambiamento.

Causal inference in integrazione multisensoriale

Un problema chiave è determinare quando segnali da diverse modalità dovrebbero essere integrati (provengono dalla stessa fonte) vs. segregati (provengono da fonti diverse).

Modello bayesiano di causal inference:

Due ipotesi:

- C=1: Segnali hanno causa comune → integrare
- C=2: Segnali hanno cause separate → segregare

Prior:

$$P(C=1) = p_{\text{common}}$$

Likelihood delle osservazioni (s_{vis} , s_{aud}) dato causa comune:

$$P(s_{\text{vis}}, s_{\text{aud}} | C=1) = N(s_{\text{vis}}; s, \sigma^2_{\text{vis}}) N(s_{\text{aud}}; s, \sigma^2_{\text{aud}})$$

Likelihood dato cause separate:

$$P(s_{\text{vis}}, s_{\text{aud}} | C=2) = N(s_{\text{vis}}; s_{\text{vis}}, \sigma^2_{\text{vis}}) N(s_{\text{aud}}; s_{\text{aud}}, \sigma^2_{\text{aud}})$$

Posterior:

$$P(C=1 | s_{\text{vis}}, s_{\text{aud}}) \propto P(s_{\text{vis}}, s_{\text{aud}} | C=1) p_{\text{common}}$$

$$P(C=2 | s_{\text{vis}}, s_{\text{aud}}) \propto P(s_{\text{vis}}, s_{\text{aud}} | C=2) (1 - p_{\text{common}})$$

Decisione di integrazione:

Se $P(C=1 | s_{\text{vis}}, s_{\text{aud}}) > 0,5 \rightarrow$ integrare

Altrimenti $\rightarrow$ segregare

Applicazione a change detection:

- Suono + cambiamento visivo congruente $\rightarrow$ alta $P(C=1) \rightarrow$ integrazione $\rightarrow$ facilitazione
- Suono incongruente $\rightarrow$ bassa $P(C=1) \rightarrow$ segregazione $\rightarrow$ nessun beneficio

Questo spiega perché audio-visual congruence è critica per facilitazione cross-modale.

Verso una comprensione quantitativa unificata

La ricerca sulla change blindness e sui limiti dell'attenzione ha rivelato che la percezione umana opera sotto severi vincoli di capacità che possono essere quantificati matematicamente. Questi vincoli derivano da principi fondamentali:

1. **Limiti di bandwidth:** Il sistema nervoso può trasmettere solo una quantità limitata di informazione per unità di tempo, creando colli di bottiglia nell'elaborazione.
2. **Costi energetici:** Rappresentazioni dense e complete sarebbero proibitivamente costose dal punto di vista metabolico.
3. **Principi di efficienza:** Sistemi sensoriali che sfruttano le regolarità statistiche dell'ambiente possono rappresentare l'informazione rilevante in modo più compatto.
4. **Trade-off precisione-generalizzazione:** Rappresentazioni troppo dettagliate possono ostacolare la generalizzazione e il riconoscimento robusto attraverso variazioni.

Le rappresentazioni sparse che emergono da questi vincoli rendono il sistema suscettibile a change blindness, ma permettono anche un'elaborazione efficiente e flessibile dell'informazione visiva. La natura frammentaria della percezione non è un difetto ma una caratteristica del design, ottimizzata per le esigenze ecologiche degli organismi viventi.

Framework unificato: Information-theoretic perspective

Possiamo tentare di unificare questi principi in un framework information-theoretic. La percezione può essere vista come un problema di compressione lossy ottimale dove il sistema deve:

Massimizzare: Informazione preservata sulla scena

$I(\text{Scene}; \text{Representation})$

Soggetto a vincoli:

1. Capacità limitata: $H(\text{Representation}) \leq C$
2. Costo energetico: $E(\text{Representation}) \leq E_{\text{max}}$
3. Latenza: $T(\text{processing}) \leq T_{\text{max}}$

La soluzione ottimale definisce una rappresentazione che:

- È sparse (per vincoli di capacità ed energia)
- Prioritizza informazione task-relevant (per massimizzare utilità)
- Usa prior statistici per compressione (per efficienza)

Change blindness emerge naturalmente: cambiamenti in informazione non preservata nella rappresentazione compressa non possono essere rilevati.

Direzioni future

Gli sviluppi futuri nella comprensione quantitativa di questi fenomeni richiederanno l'integrazione di misurazioni comportamentali, neurali e computazionali, permettendo modelli sempre più precisi che possano predire non solo cosa vediamo, ma anche, e forse più significativamente, cosa non vediamo.

Approcci promettenti:

1. **Neuroimaging ad alta risoluzione:** 7T fMRI, MEG source localization per tracciare rappresentazioni neurali con precisione senza precedenti
2. **Large-scale neural recordings:** Neuropixels e tecnologie simili permettono registrazione simultanea da migliaia di neuroni, rivelando dinamiche di popolazione
3. **Closed-loop experiments:** Sistemi che modificano stimoli in tempo reale basandosi su stato neurale istantaneo
4. **Modelli generativi profondi:** VAEs, GANs, diffusion models come modelli computazionali di rappresentazioni interne
5. **Integrazione teoria-dati:** Metodi bayesiani per inferire modelli cognitivi da dati multi-source (comportamento + fMRI + EEG + eye-tracking)

Questa comprensione matematica della incompletezza percettiva rappresenta uno dei contributi più profondi delle neuroscienze cognitive moderne alla nostra comprensione della mente e della coscienza.

La realtà esperienziale non è un'immagine fedele del mondo esterno, ma una ricostruzione attivamente generata, basata su frammenti di informazione sensoriale e guidata da aspettative e conoscenza precedente. Questa natura costruttiva della percezione, resa evidente dalla change blindness, ci ricorda che la coscienza è un'inferenza, non un riflesso, e che i limiti di questa inferenza possono essere compresi attraverso l'applicazione rigorosa di principi matematici e computazionali.

La matematica di ciò che non vediamo illumina tanto sulla natura della mente quanto la matematica di ciò che vediamo, rivelando l'eleganza dei compromessi evoluti che permettono ai sistemi cognitivi di navigare un mondo di complessità infinita con risorse finite.